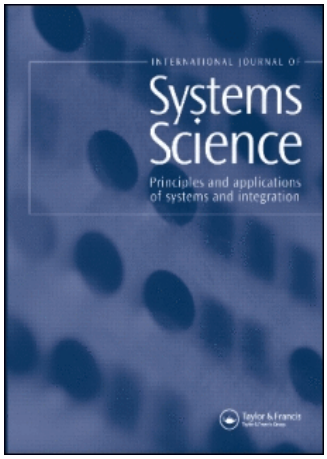

International Journal of Systems Science

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713697751

Model selection approaches for non-linear system identification: a review

X. Hong [a]; R. J. Mitchell [a]; S. Chen [b]; C. J. Harris [b]; K. Li [c]; G. W. Irwin [c]
[a] School of Systems Engineering, University of Reading, UK
[b] School of Electronics and Computer Science, University of Southampton, Southampton, UK
[c] School of Electronics, Electrical Engineering and Computer Science Queen's University Belfast, UK

Online Publication Date: 01 October 2008

To cite this Article: Hong, X., Mitchell, R. J., Chen, S., Harris, C. J., Li, K. and Irwin, G. W. (2008) 'Model selection approaches for non-linear system identification: a review', International Journal of Systems Science, 39:10, 925 — 946

To link to this article: DOI: 10.1080/00207720802083018
URL: http://dx.doi.org/10.1080/00207720802083018

International Journal of Systems Science
Vol. 39, No. 10, October 2008, 925–946

Model selection approaches for non-linear system identification: a review

X. Hong[a]*, R.J. Mitchell[a], S. Chen[b], C.J. Harris[b], K. Li[c] and G.W. Irwin[c]

[a]*School of Systems Engineering, University of Reading, Reading, UK;* [b]*School of Electronics and Computer Science, University of Southampton, Southampton, UK;* [c]*School of Electronics, Electrical Engineering and Computer Science Queen's University Belfast, Belfast, UK*

(*Received 28 November 2006; final version received 6 August 2007*)

The identification of non-linear systems using only observed finite datasets has become a mature research area over the last two decades. A class of linear-in-the-parameter models with universal approximation capabilities have been intensively studied and widely used due to the availability of many linear-learning algorithms and their inherent convergence conditions. This article presents a systematic overview of basic research on model selection approaches for linear-in-the-parameter models. One of the fundamental problems in non-linear system identification is to find the minimal model with the best model generalisation performance from observational data only. The important concepts in achieving good model generalisation used in various non-linear system-identification algorithms are first reviewed, including Bayesian parameter regularisation and models selective criteria based on the cross validation and experimental design. A significant advance in machine learning has been the development of the support vector machine as a means for identifying kernel models based on the structural risk minimisation principle. The developments on the convex optimisation-based model construction algorithms including the support vector regression algorithms are outlined. Input selection algorithms and on-line system identification algorithms are also included in this review. Finally, some industrial applications of non-linear models are discussed.

Keywords: adaptive learning; cross validation; model selection; model generalisation; system identification; control engineering

1. Introduction

System identification, as a subject of control engineering, refers to the procedure of building a mathematical description of the dynamic behaviour of a system/process from measured data so as to provide accurate prediction of the future behaviour for given inputs (Eykhoff 1974; Goodwin and Sin 1984; Ljung 1987; Aström and Wittenmark 1989; Söderström and Stoica 1989). The two major sub-problems in system identification are (1) to determine the model structure describing the functional relationship between the system input and output variables; and (2) to estimate the model parameters that specify any model within a chosen or derived model structure. The initial and natural approach to system identification from sequential data was to use linear difference equations between input and output observations. A number of mature linear-system estimation theories based on time series have been established over the past 40 years (Aström and Eyhhoff 1971; Box and Jenkins 1976; Priestley 1981), including adaptive methods that infer system parameters on-line using recursive estimation (Ljung and Söderström 1983; Young 1984). Whilst most early research has focussed on linear-time invariant systems, recent linear identification research has considered the identification of continuous systems (Rao 2006), subspace identification methods (Goethals, Van Gestel, Suykens, Van Dooren, and De Moor 2003; Markovsky, Willems, Rapisaida, and de Moor 2005) and errors-in-the-variable methods (Söderström 2006), etc.

A primary measure of model quality is its approximation accuracy to the unknown underlying process. Since most practical systems are non-linear to some extent, non-linear models are often required to achieve acceptable modelling performance. Defining the input of a non-linear discrete system as $u(t)$, the system output as $y(t)$, and given a training dataset D_N consisting of N input/output data pairs $\{u(t), y(t)\}_{t=1}^{N}$, the fundamental goal is then to find

$$y(t) = f(\mathbf{x}(t), \boldsymbol{\theta}) + e(t) \tag{1}$$

where the underlying function $f(\cdot)$ is unknown, $\boldsymbol{\theta}$ is the vector of associated parameter and $e(t)$ is the noise, which is often assumed to be independent and identically distributed (i.i.d.) with constant

*Corresponding author. Email: x.hong@reading.ac.uk

ISSN 0020–7721 print/ISSN 1464–5319 online

DOI: 10.1080/00207720802083018
http://www.informaworld.com

variance σ^2. The model input vector is formed using $\mathbf{x}(t) = [y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u), e(t-1), \ldots, e(t-n_e)]^T$, and n_y, n_u and n_e are lags of past output, input and noise used in constructing the model. Equation (1) leads to the non-linear auto-regressive moving average with exogeneous input (NARMAX) model representation (Leontaritis and Billings 1985), which provides a unified representation for a wide class of non-linear systems.

Supposing that the true underlying function $f(\cdot)$ is continuous and smooth (in most engineering systems this assumption is generally satisfied), the problem of non-linear system identification can be regarded as a functional approximation problem that finds an estimator $\hat{f}$ of f. Many types of non-linear models may be chosen by the users in order to approximate f (Ljung and Vicino 2005; Söderström, Van den Hof, Wahlberg, and Weiland 2005). Various modelling paradigms have been investigated, e.g. piecewise linear models (Billings and Voon 1987), rational polynomial models (Mao, Billings, and Zhu 1999), Hammerstein/WienerHammerstein models (Greblicki 1989; Bai 1998, 2004), projection pursuit regression (PPR) and multivariate adaptive regression splines (MARS) (Friedman and Stuetzle 1981; Friedman 1991), Gaussian processes (Neal 1996; Mackay 1997) and recurrent neural networks (Mandic and Chambers 2001). The multilayer perceptron (MLP) (Funahashi 1989) and radial basis function (RBF) neural networks have been proven to be capable of representing a class of unknown non-linear input–output mappings with arbitrary small approximation error capability (Powell 1985; Broomhead and Lowe 1988; Cybenko 1989; Hornik, Stinchcombe, and White 1989; Girosi and Poggio 1990; Park, EJ-Sharkawi, and Marks 1991). In approximation theory, a general way of representing functions is via a linear regression of non-linear basis functions. Many existing approximation schemes in the field of approximation theory have been naturally adopted into the neural networks family such as B-spline neural network of (Kavli 1993; Brown and Harris 1994; Harris, Hong, and Gan 2002), wavelets (Zhang 1993; Sjöberg et al. 1995; Juditsky et al. 1995), or more generally the generalised single hidden layer neural nets (Adeney and Korenberg 2000; Huang, Zhu, and Siew 2004b). These models represent non-linear input/output relationship with a linear-in-the-parameters structure given by

$$\hat{f}(\mathbf{x}(t), \boldsymbol{\theta}) = \sum_{i=1}^{m} \phi_i(\mathbf{x}(t))\theta_i \tag{2}$$

where $\phi_i(\mathbf{x}(t))$ is a known non-linear basis function mapping, such as RBF, polynomial or B-spline functions, θ_i are unknown parameter and m is the number of basis functions in the model.

The linear-in-the-parameter models are well structured for adaptive learning, have provable learning and convergence conditions, have the capability of parallel processing and have clear applications in control engineering (Murray-Smith and Johansen 1997; Fabri and Kadirkamanathan 2001; Ruano 2005). There are still, however, some major challenges and obstacles in non-linear system identification:

Model generalisation: The ultimate objective should be to produce a model that captures the true underlying dynamics and predicts accurately the output for unseen data. The model identified using a finite training dataset should not just have good accuracy over the training dataset but also be tested on an independent dataset. As f is unknown, $y(t)$ is used as the target for training model $\hat{f}$ and the modelling accuracy of $\hat{f}(t)$ to the target $y(t)$ is increasing as model complexity increases. Thus, overfitting to the noise contained in $y(t)$ may occur if model accuracy is over pursued. How to define and achieve model generalisation is central to all learning machines.

Model interpretation: A model is often used to interpret the properties of the process that it represents and to extract the knowledge of the underlying system. Many good properties of linear systems do not hold for non-linear models, e.g. the free exchangeability of model representations between the time domain and frequency domain. The parameters in a linear model can often be associated with the physical nature of the system. Due to the non-linear nature and higher model complexity, it is very difficult for non-linear models to be utilised for explanation of the structural characteristics of the system, unless there is a *deliberate* effort in revealing information in the modelling strategy/process by the modeller.

The curse of dimensionality: The number of model parameters can easily be excessive in relation to the size of the dataset in the construction of model (2), e.g. the basis functions may either be built up using the tensor products from univariate functions (as in B-spline networks) or radial construction from training data (as in RBF neural networks). An over-parameterised model is ill-conditioned, by which the parameters cannot be estimated with sufficient accuracy, leading to poor model generalisation performance. Note that in linear models persistent excitation (PE) is generally achieved through input signal design (Ljung 1987; Söderström and Stoica 1989) in order to guarantee the non-singularity of the regression matrix spanning the input space. In the case of non-linear models, both input signal and regression matrix design are required

so as to ensure the non-singularity of regression matrix for a non-linear model due to non-linearities.

Computational complexity: Associated with the curse of dimensionality is an excessively high computational complexity. Non-linear system identification is inherently an intractable problem. For practical applications, efficient non-linear system-identification algorithms are highly desirable. The algorithm design demands innovative computational engineering using an interaction between system theory, statistics, optimisation theory, intelligent learning and linear algebra.

Input selection: For many non-linear models, the size of model can increase exponentially fast as the model input dimension ($\mathbf{x}$) increases. Using too many input terms may have undesirable effects on the modelling performance through either incorrect input setting or overparametrisation. For the actual system output, some input variables may be redundant or would become insignificant if some other input variables were present in the model. Input selection as a preprocessing procedure can significantly help improve network performance and model interpretation. However, optimal input selection is often an intractable task, and efficient input selection algorithm is always an important element in many pattern recognition applications.

Robustness and noise rejection: For linear system identification, there exist effective techniques to achieve robust estimation and noise rejection. Conventional linear system identification is based on the assumptions of linear-time invariance of the process, usually with additive Gaussian noise. Yet, the majority of real dynamics processes are complex, non-linear, non-stationary, stochastic and partially unknown. Conventional-learning algorithms often have limitations when applied to the real system processes. For improved performance on model robustness and noise rejection, it is necessary to investigate algorithms that deal with processes, which are both non-linear and non-Gaussian.

On-line non-linear system identification: In many applications, the models are to be inferred for real time operation where the data samples are available sequentially. On-line system-identification algorithms are computationally advantageous in that the model is updated following the arrival of new data rather than being relearnt from scratch. The concept of on-line learning is also an important concept in intelligent systems as the natural human learning behaviour is to build up *a posteriori* knowledge based upon *a priori* knowledge. Although the linear recursive-identification algorithms can update model parameters for a fixed model structure, this may be limited whenever there is a need to update the model structure. On-line estimation algorithms based on variable structure and sparse models with fast computation ability is an important current area of research.

Whilst there is an abundance of publications in the area of the non-linear system identification, this survey emphasises the concepts and computational techniques of non-linear system identification for a class of linear-in-the-parameter models, which have universal approximation capabilities. In the following, the important concepts in various non-linear system-identification algorithms for such models are reviewed. Developments in both stepwise-selection algorithms and convex optimisation-based model-construction algorithms including support vector regression (SVR) algorithms are then outlined. Next the input selection algorithms and on-line system identification algorithms are briefly reviewed. Finally, industrial applications of the linear-in-the-parameter models are briefly surveyed.

2. Model generalisation

2.1 *Parameter regularisation*

The major purpose of model construction is to produce good generalisation (the capability to provide approximation to the true system output for new input data). The technique of parameter regularisation is one of the primary tools for improving model generalisation. Note that the effects of model parameter estimation on model generalisation can be analysed via the mean square error (MSE) of a parameter estimator, which can be used as a measure of model generalisation. Suppose that the true dynamics of a system can be represented by model (2) and parameterised with an estimator $\widehat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m]^T$. The MSE of $\widehat{\boldsymbol{\theta}}$ of (2) is given by

$$\begin{aligned} &E[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})] \\ &\quad = E[(\widehat{\boldsymbol{\theta}} - E(\widehat{\boldsymbol{\theta}}) + E(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}} - E(\widehat{\boldsymbol{\theta}}) + E(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta})] \\ &\quad = \mathrm{Var}(\widehat{\boldsymbol{\theta}}) + \mathrm{bias}(\widehat{\boldsymbol{\theta}})^2. \end{aligned} \tag{3}$$

where $\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = E[(\widehat{\boldsymbol{\theta}} - E(\widehat{\boldsymbol{\theta}}))^T(\widehat{\boldsymbol{\theta}} - E(\widehat{\boldsymbol{\theta}}))]$ and $\mathrm{bias}(\widehat{\boldsymbol{\theta}}) = E(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$.

It is known in statistics that in general ($m > 3$), there exists some biased parameter estimator to dominate any unbiased parameter estimator in terms of the MSE (James and Stein 1961). One way of improving model generalisation is therefore to design the bias that the model variance can be reduced significantly at the cost of a small bias as to achieve a good bias/variance trade-off.

The method of regularisation or ridge regression is a simple, yet effective way of achieving a good bias/variance trade-off (Hoerl and Kennard 1970; Marquardt 1970). For (2), over the training dataset D_N, the regression matrix $\mathbf{\Phi}$ is arranged as

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1(\mathbf{x}(1)) & \phi_2(\mathbf{x}(1)) & \cdots & \phi_m(\mathbf{x}(1)) \\ \phi_1(\mathbf{x}(2)) & \phi_2(\mathbf{x}(2)) & \cdots & \phi_m(\mathbf{x}(2)) \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \phi_1(\mathbf{x}(N)) & \phi_2(\mathbf{x}(N)) & \cdots & \phi_m(\mathbf{x}(N)) \end{bmatrix}.$$

and $\mathbf{y} = [y(1), \ldots, y(N)]^T$.

For illustration, let $e(t) \sim N(0, \sigma^2)$, and the least squares estimator $\widehat{\boldsymbol{\theta}}_{LS}$ be obtained as $\widehat{\boldsymbol{\theta}}_{LS} = [\mathbf{\Phi}^T\mathbf{\Phi}]^{-1}\mathbf{\Phi}^T\mathbf{y}$ via minimising cost $J = [\mathbf{y} - \mathbf{\Phi}\boldsymbol{\theta}]^T \times [\mathbf{y} - \mathbf{\Phi}\boldsymbol{\theta}]$, which is also the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$. We have $\widehat{\boldsymbol{\theta}}_{LS} \sim N(\boldsymbol{\theta}, [\mathbf{\Phi}^T\mathbf{\Phi}]^{-1}\sigma^2)$. The MSE of the parameter estimator $\widehat{\boldsymbol{\theta}}_{LS}$, is given by

$$\begin{aligned} E[(\widehat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta})] &= \text{trace}[\text{cov}(\widehat{\boldsymbol{\theta}}_{LS})] \\ &= \sigma^2 \text{trace}([\mathbf{\Phi}^T\mathbf{\Phi}]^{-1}) \\ &= \sigma^2 \sum_{i=1}^{m} \frac{1}{\lambda_i} \end{aligned} \tag{4}$$

where λ_i are the eigenvalues of the positive definite matrix $\mathbf{\Phi}^T\mathbf{\Phi}$, which are assumed to be in the order $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m = \lambda_{\min} > 0$. Note that in the case of ill-conditioning, $\lambda_{\min}$ can be very small so that the MSE of $\widehat{\boldsymbol{\theta}}_{LS}$ is very large. Ridge regression is a biased estimator $\widehat{\boldsymbol{\theta}}_R$ (Hoerl and Kennard 1970; Marquardt 1970) and is obtained by minimising $J_R = [\mathbf{y} - \mathbf{\Phi}\boldsymbol{\theta}]^T[\mathbf{y} - \mathbf{\Phi}\boldsymbol{\theta}] + \mu\boldsymbol{\theta}^T\boldsymbol{\theta}$ and given by

$$\widehat{\boldsymbol{\theta}}_R = [\mathbf{\Phi}^T\mathbf{\Phi} + \mu\mathbf{I}]^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{5}$$

where $\mu > 0$ is called the regularisation parameter. It is shown (Hoerl and Kennard 1970) that

$$\begin{aligned} E[(\widehat{\boldsymbol{\theta}}_R - \boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}}_R - \boldsymbol{\theta})] = \sigma^2 &\sum_{i=1}^{m} \frac{\lambda_i}{(\lambda_i + \mu)^2} \\ &+ \mu^2\boldsymbol{\theta}^T(\mathbf{\Phi}^T\mathbf{\Phi} + \mu\mathbf{I})^{-2}\boldsymbol{\theta} \end{aligned} \tag{6}$$

As μ increases, the first term of (6) (the variance) is monotonically decreasing whilst the second term of (6) (the bias) is monotonically increasing, and there always exists a $\mu > 0$ such that $E[(\widehat{\boldsymbol{\theta}}_R - \boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}}_R - \boldsymbol{\theta})] < E[(\widehat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta})^T(\widehat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta})]$.

The regularised parameter estimator $\widehat{\boldsymbol{\theta}}_R$ obtained by optimising J_R is equivalent to the maximum *a posteriori* pdf (MAP) of parameters in a Bayesian approach (MacKay 1991). By Bayesian Theorem

$$p(\boldsymbol{\theta}|D_N, \alpha, \beta) = \frac{p(\boldsymbol{\theta}|\alpha)p(D_N|\boldsymbol{\theta}, \beta)}{p(D_N|\alpha, \beta)} \tag{7}$$

where $p(\boldsymbol{\theta}|\alpha)$ is the prior, $p(D_N|\boldsymbol{\theta}, \beta)$ the likelihood and $p(D_N|\alpha, \beta)$ the evidence that does not dependent on θ explicitly.

Assuming that the observations are independent, so

$$p(D_N|\boldsymbol{\theta}, \beta) = \frac{\exp\left[-(\beta/2)\sum_{t=1}^{N}\left[y(t) - \sum_{i=1}^{m}\phi_i(\mathbf{x}(t))\theta_i\right]^2\right]}{Z_D(\beta)} \tag{8}$$

where $\beta = (1/\sigma^2)$ and $Z_D(\beta) = (2\pi/\beta)^{N/2}$ is a normalising coefficient. By using prior knowledge of $p(\boldsymbol{\theta}|\alpha)$ that controls superfluous parameters, the MAP estimator is a solution that resolves the inadequacy of ML estimator for improved generalisation. If the prior $p(\boldsymbol{\theta}|\alpha)$ for the parameters is Gaussian

$$p(\boldsymbol{\theta}|\alpha) = \exp\left(-\frac{\alpha}{2}\sum_{i=1}^{m}\theta_i^2\right) \Big/ Z_{\boldsymbol{\theta}}(\alpha) \tag{9}$$

where $Z_{\boldsymbol{\theta}}(\alpha) = (2\pi/\alpha)^{m/2}$ is a normalising coefficient, the MAP estimator is equivalent to the regularised least squares parameter estimator via minimising J_R (MacKay 1991; Chen 2002) with the regularisation parameter $\mu = (\alpha/\beta)$.

A further question is how to determine the regularisation parameter through some optimisation procedure. The optimal value of μ through the generalised cross validation (*GCV*) (Section 2.2) has been developed by Wahba (1990) and Hastie and Tibshirani (1996). Alternatively in the Bayesian's formulation, this is found through an evidence maximisation procedure (MacKay 1991).

The above parameter regularisation, based on l_2 norm regulariser in the objective function, is closely related to the structure risk principle in Section 4.1, and l_1 regularisation in Section 4.3. A promising approach for non-linear system identification, the least squares support vector machine (LSSVM) (Suykens, Van Gestel, De Branbanter, De Moor, and Vandewalle 2002; Goethals, Pelckmans, Suykens, and De Moor 2005; Espinoza, Suykens, and De Moor 2005b), is also an application of l_2 norm parameter regularisation in the kernel feature space (Section 4).

System identification is simply an optimisation problem. The use of different objective functions, often in the form of a trade-off between the model fit and structural/parametric constraints, leads to alternative models. The model selective criteria are used for the the discrimination of the model's generalisation capability amongst different models.

2.2 *Model selective criteria*

2.2.1 *Model selective criteria based on cross validation*

Information theoretic metrics of a model's generalisation capability are of great importance in statistical learning including non-linear system identification. A fundamental concept in the evaluation of model generalisation capability is that of cross validation (Stone 1974), which is often used to derive the information theoretic metrics. Model selective criteria can be used for either predicting a model's performance on unseen data or evaluating a model's quality amongst other competitive models. Suppose that a system modelled by (2) is parameterised with $\widehat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m]^T$ and let $\widehat{\sigma}^2$ be the estimator of σ^2. The sum of squared errors over the estimation dataset is given by $SSE(\widehat{\boldsymbol{\theta}}) = \sum_{t=1}^{N}[y(t) - \sum_{i=1}^{m} \phi_i(\mathbf{x}(t))\hat{\theta}_i]^2$. Various model selection criteria have been introduced such as the *GCV* (as detailed below), Mallow's C_p, final prediction error (*FPE*), Akaike's information criteria (*AIC*) and the predicted residual error sum of squares (*PRESS*) statistic (Allen 1974; Akaike 1974; Barron 1984; Miller 1990; Moody 1994) given, respectively, by

$$\begin{aligned} C_p &= \frac{SSE(\widehat{\boldsymbol{\theta}})}{\widehat{\sigma}^2} + 2m - N \\ FPE &= \frac{SSE(\widehat{\boldsymbol{\theta}})}{N} \cdot \left(\frac{N+m}{N-m}\right) \\ AIC &= N \log\left(\frac{SSE(\widehat{\boldsymbol{\theta}})}{N}\right) + 2m \\ PRESS &= \widehat{\sigma}^2\left(1 + \frac{2m}{N}\right) \end{aligned} \tag{10}$$

In order to illustrate how these model criteria are obtained, consider one commonly used version of cross validation, the leave one out (LOO) cross validation (Stone 1974), The idea is that, for any predictor, each data point in the estimation dataset D_N is sequentially set aside in turn, a model is then estimated using the remaining $(N-1)$ data, and the prediction error is derived for the data point that was removed. For convenience, $\hat{y}^{(-)}(i)$ is defined as the output for input $\mathbf{x}(i)$, of model (1) estimated using the LOO, dataset, $D_N \backslash \{\mathbf{x}(i), y(i)\}$. The LOO errors corresponding to $\hat{y}^{(-)}(i)$ are given by (Stone 1974)

$$\zeta^{(-)}(i) = y(i) - \hat{y}^{(-)}(i) \tag{11}$$

The mean squares of LOO errors $E[(\zeta^{(-)}(i))^2]$ is often used as the metric of the model generalisation errors. The LOO errors can be calculated without actually splitting the dataset for the linear-in-the-parameters model of (2) as parameterised with the least squares estimator. Denoting the model residual sequence as $\zeta(i) = y(i) - \sum_{j=1}^{m} \phi_j(\mathbf{x}(i))\hat{\theta}_j$, it is shown (Stone 1974) that

$$\zeta^{(-)}(i) = \frac{\zeta(i)}{1 - \phi(i)[\boldsymbol{\Phi}^T\boldsymbol{\Phi}]^{-1}[\phi(i)]^T} \quad i = 1, \ldots, N \tag{12}$$

where $\phi(i)$ denotes the ith row vector of $\boldsymbol{\Phi}$. Similarly, for the linear-in-the-parameters model as parameterised with the regularised least squares estimator of (5), the LOO errors are (Wahba 1990; Green and Silverman 1994)

$$\zeta^{(-)}(i) = \frac{\zeta(i)}{1 - \phi(i)[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mu\mathbf{I}]^{-1}[\phi(i)]^T} \quad i = 1, \ldots, N \tag{13}$$

Based on the lower triangular, diagonal, lower triangular (LDL) matrix decomposition of a positive definite matrix, a computationally efficient algorithm calculating LOO errors without matrix inversion is available (Green and Silverman 1994). A popular variant of the mean squares of LOO errors is to replace the denominator in (13) for all data samples by their average value, leading to the *PRESS* statistic (see (10)) (Allen 1974; Miller 1990)

$$\begin{aligned} PRESS &= \frac{\sum_{i=1}^{N} \zeta^2(i)}{N(1 - \mathrm{tr}[\mathbf{H}]/N)^2} \\ &\approx \widehat{\sigma}^2\left(1 + \frac{2m}{N}\right) \end{aligned} \tag{14}$$

where $\mathbf{H} = \boldsymbol{\Phi}[\boldsymbol{\Phi}^T\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}^T$, $\widehat{\sigma}^2 = (1/N)\sum_{i=1}^{N} \zeta^2(i)$ under the assumption that m is much smaller than the number of data samples N. Let $\mathbf{H}(\mu) = \boldsymbol{\Phi}[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mu\mathbf{I}]^{-1}\boldsymbol{\Phi}^T$, the *GCV* is given by

$$GCV(\mu) = \frac{\sum_{i=1}^{N} \zeta^2(i)}{N(1 - \mathrm{tr}[\mathbf{H}(\mu)]/N)^2} \tag{15}$$

which corresponds to (13) for the regularised least squares estimator. The *GCV* can be optimised with respect to μ to find the optimal μ (Wahba 1990).

The sensitivity of the model selective criteria to a unit increase in the number of model parameters can be utilised to reveal the differences of the set of model selective criteria especially for small datasets and small model size (Bossley 1997). Despite the differences amongst the set of model selective criteria, there are common characteristics in which the above criteria are equivalently asymptotic under general conditions (Nishii 1984), and these share the form of trade-off between the goodness of fit in the estimation data and the model complexity. Consequently, information theoretic metrics are consistent with the basic principle of parsimony of using the smallest possible model (Occam's razor).

A more sophisticated perspective is the bias/variance dilemma (Geman, Bienenstock, Dowsat and 1992; Moody 1994) in which a simple model tends to generalise well when it is used to infer the true system/process dynamics based on a finite dataset. The bias/variance dilemma simply states that the model generalisation error can be decomposed into two components, the bias and the variance. A model with high approximation capability has high variability and may fit the estimation dataset too well rather than generalise for a new dataset. The bias refers to any flexibility constraints on the model. A smaller model with less approximation capacity has higher bias.

2.2.2 *Model selective criteria based on experimental design*

The optimum experimental design is a subject of statistics (Atkinson and Donev 1992; Myers and Montgomery 1995) used to construct smooth network response surfaces based on the setting of the experimental variables under well-controlled experimental conditions. In optimum design, model adequacy is evaluated by design criteria, which are statistical measures of goodness of experimental designs by virtue of design efficiency and experimental effort. Two examples of experimental design in linear system identification are the well-known PE condition for input signal design (Ljung 1987; Söderström and Stoica 1989) and the more recently proposed robust input spectral density design (Rojas, Welsh, Goodwin, and Fever 2007). A recent review in the context of linear system identification for control can be found (Gevers 2005). Despite the fact that the design of experiments for general non-linear system identification lacks a formal coverage, the underlying idea is closely related to the active learning (Plutowski and White 1993; Cohn, Ghahramani, and Jordan 1996), the paradigm of local modelling (Murray-Smith 1994) and the experiment design combining with a real-time control experiment (Stewart, Fleming, and Mackenzie, 2003). These methods attack the problem of the curse of dimensionality by acquiring datasets or defining model structures that are relevant to the operating region/tasks (Murray-Smith 1994; Cohn et al. 1996; Stewart et al. 2003).

One way of applying the concept of optimal experiment design for model selection is to measure the model adequacy as a function of the eigenvalues of the design matrix, $\mathbf{\Phi}^T\mathbf{\Phi}$. It is well known that a model based on least squares estimates tends to be unsatisfactory for a near ill-conditioned regression matrix (or design matrix). Note that the MSE of $\widehat{\boldsymbol{\theta}}_{LS}$ given by (4) is very large when $\lambda_{\min}$ is close to zero. It is natural to consider model subset selection in the framework of the optimal experiment design. The subset model is constructed from the full model with regression matrix $\mathbf{\Phi}$ by using n_θ regressors selected from m regressors in $\mathbf{\Phi}$, $n_\theta \ll m$. Define the resultant regression matrix as $\mathbf{\Phi}_k \in \Re^{N\times n\theta}$, and the resultant design matrix by $\mathbf{\Phi}_k^T\mathbf{\Phi}_k$, and λ_k, $k=1,\ldots,n_\theta$ are still used to denote the eigenvalues of $\mathbf{\Phi}_k^T\mathbf{\Phi}_k$ for simplicity. The following two experimental design criteria in the context of model subset selection may be used for subset selection.

Definition 1: The A-optimality criterion minimises the sum of the variance of a parameter estimate vector $\hat{\boldsymbol{\theta}}_{LS}$

$$\min\left\{J_A = \mathrm{tr}\Big[\mathrm{cov}(\hat{\boldsymbol{\theta}}_{LS})\Big] = \sigma^2\sum_{k=1}^{n_\theta}\frac{1}{\lambda_k}\right\}. \tag{16}$$

Definition 2: The D-optimality criterion maximises the determinant of the design matrix of $\mathbf{\Phi}_k^T\mathbf{\Phi}_k$

$$\max\left\{J_D = \det(\mathbf{\Phi}_k^T\mathbf{\Phi}_k) = \prod_{k=1}^{n_\theta}\lambda_k\right\}. \tag{17}$$

These criteria (Atkinson and Donev 1992) inherently improve model robustness by favouring models with smaller condition numbers to ensure a low value of MSE for $\widehat{\boldsymbol{\theta}}_{LS}$. It should be noted that, when used for model subset selection, these criteria are not related to the significance of the regressors in explaining the output variable, i.e. the model approximation capability of the final model is not taken into account.

2.2.3 *Correlation-based model validation*

In classical linear model identification, correlation function-based model validation tests have been widely applied to validate the estimated models (Bohlin 1971; Box and Jenkins 1976; Söderström and Stoica 1990). Model validation tests are procedures to detect the inadequacy of the model. If the model structure and the estimated parameters are appropriate, then the residual sequence $\zeta(t) = y(t) - \sum_{i=1}^{m}\phi_i(\mathbf{x}(t))\hat{\theta}_i$ should be unpredictable from all linear and non-linear combinations of the past inputs $u(t)$ and outputs $y(t)$. Despite the complexity and difficulties in designing correlation function-based model validation tests, there are various correlation-based model validation tests developed to deal with the effects of system non-linearity (Mao and Billings 2000; Zhang, Zhu, and Longden 2007).

3. Model construction using stepwise selection algorithms

A practical non-linear modelling principle is to find the smallest model that generalises well. Sparse models are preferable in engineering applications, since a model's computational complexity scales with its model complexity. Moreover, a sparse model is easier to interpret from the viewpoint of knowledge extraction. It is important to acknowledge that the non-linear system identification is an intractable optimisation problem as any algorithm can only aim to solve the problem within some hypothesis space. For practical applications, optimisation algorithms with computational simplicity are highly desirable. In any practical non-linear system-identification algorithm, the problems are initially formulated as some tractable problems of which suboptimal solutions can be obtained. Considering the subset selection of choosing n_θ from m candidate terms and taking $m=500$ and $n_\theta=40$, there are $m!/(n_\theta!(m-n_\theta)!)=2.2443\times10^{59}$ possible model structures to select from. Forward/backward subset selection algorithms are greedy algorithms that aim to optimise some objective function at each regression stage. The classical forward (backward) approach appends (removes) a model regressor one at a time based on largest improvement (least deterioration) in model fit (Miller 1990). In the forward subset selection of choosing n_θ from m candidate terms, for the same $m=500$ and $n_\theta=40$, the number of candidate model evaluation is reduced to $\sum_{k=0}^{n_\theta}(m-k)<n_\theta m=2\times10^4$.

Among various stepwise subset selection algorithms, the forward orthogonal least squares (OLS) is an efficient non-linear system-identification algorithm (Korenberg 1988; Chen, Billins, and Luo 1989) which selects regressors in a forward manner by virtue of their contribution to the maximisation of the model error reduction ratio (ERR). The forward OLS estimator involves selecting a set of n_θ variables $\phi_k=[\phi_k(1),\ldots,\phi_k(N)]^T$, $k=1,\ldots,n_\theta$, from m regressors to form a set of orthogonal basis $\mathbf{p}_k$, $k=1,\ldots,n_\theta$, in a forward regression manner. To produce a model with good generalisation capabilities, model selection criteria such as the *AIC* (Akaike 1974) are usually incorporated into the procedure to determinate the model construction process. The OLS algorithm has become a popular modelling tool for the associative neural networks such as fuzzy/neurofuzzy systems (Wang and Mendel 1992; Hong and Harris 2001a), wavelets neural networks (Zhang 1993; Billings and Wei 2005). The algorithm has also been utilised in a wide range of engineering applications, e.g. aircraft gas turbine modelling (Chiras, Evans, and Rees 2001), fuzzy control of MIMO non-linear systems (Gao and Er 2003), power system control (Tsang and Chan 2005) and fault detection (Luh and Cheng 2004).

3.1 *The locally regularised orthogonal least squares algorithm*

Regularisation techniques have been incorporated into the forward selection (Orr 1995) and a regularised orthogonal least squares (ROLS) algorithm has been introduced to reduce the variance of parameter estimates (Chen, Wu, and Luk 1999; Chen 2002). The advantage of ROLS is that the parameter regularisation is applied to the auxiliary parameters in orthogonal space, simplifying the calculation for parameter estimation significantly. The locally regularised orthogonal least squares (LROLS) procedure as outlined below (Chen 2002) can automatically select a subset of n_θ regressors to construct a parsimonious model.

An orthogonal decomposition of $\mathbf{\Phi}$ is

$$\mathbf{\Phi}=\mathbf{PA} \tag{18}$$

where $\mathbf{A}=\{a_{ij}\}$ is an $m\times m$ unit upper triangular matrix and $\mathbf{P}=[\mathbf{p}_1,\ldots,\mathbf{p}_m]$ is an $N\times m$ matrix with orthogonal columns that satisfy

$$\mathbf{P}^T\mathbf{P}=\mathrm{diag}\{\kappa_1,\ldots,\kappa_m\} \tag{19}$$

with

$$\kappa_k=\mathbf{p}_k^T\mathbf{p}_k,\quad k=1,\ldots,m \tag{20}$$

so that (1) and (2) can be expressed in vector form as

$$\mathbf{y}=(\mathbf{\Phi A}^{-1})(\mathbf{A}\Theta)+\boldsymbol{\zeta}=\mathbf{Pg}+\boldsymbol{\zeta} \tag{21}$$

where $\mathbf{g}=[g_1,\ldots,g_m]^T$ is an auxiliary vector. The LROLS algorithm uses the following error criterion for parameter estimation:

$$J_R=\boldsymbol{\zeta}^T\boldsymbol{\zeta}+\mathbf{g}^T\mathcal{U}\mathbf{g}. \tag{22}$$

where $\mathcal{U}=\mathrm{diag}\{\mu_1,\ldots,\mu_k,\ldots,\mu_m\}$, with μ_k's are positive regularisation parameters. Because $\zeta(t)$ is uncorrelated with regressors, it may be shown (Chen et al. 1989) that

$$g_k=\frac{\mathbf{w}_k^T\mathbf{y}}{\kappa_k+\mu_k},\quad k=1,\ldots,m. \tag{23}$$

The original model coefficient vector $\boldsymbol{\theta}=[\theta_1,\ldots,\theta_m]^T$ can then be calculated from $\mathbf{A}\boldsymbol{\theta}=\mathbf{g}$ through backsubstitution.

At the kth selection, a candidate regressor is selected as the kth basis of the subset if it produces the largest value of $[ERR]_k=(g_k^2\kappa_k/\mathbf{y}^T\mathbf{y}^T)$ from the remaining $(m-k+1)$ candidates. Equivalently, this procedure can be expressed as

$$J^{(k)}=J^{(k-1)}-\frac{1}{N}g_k^2\kappa_k \tag{24}$$

where $J^{(0)} = \mathbf{y}^T\mathbf{y}$. At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$. A possible disadvantage of using (24) for model selection is that this is not directly derived by optimising model generalisation. Note that *AIC* or other information-based criteria are usually simplified measures derived as approximation formulas that are particularly sensitive to model complexity. The model selective criteria given by (10), aside from affecting the stopping point of the model selection, does not have more discriminative power about model generalisation than using (24). This means the regressors that might cause poor model performance, e.g. too large parameter variance or ill-conditioning of the regression matrix, are not directly penalised during the model selection.

3.2 *Improvements on the model selective criteria in forward selection*

Recently, some variants of OLS algorithms have been introduced by modifying the model selective criteria based on the experimental design criteria, in order to ensure that the best model in terms of the experimental design criteria is found amongst the candidate models (Hong and Harris 2001a, 2003; Chen, Hong, and Hauis 2003a). Specifically instead of (24), the OLS algorithm with A-optimality and D-optimality algorithms use the following model selective criteria

$$\begin{aligned}
&\text{A-optimality} + \text{OLS}: \; J^{(k)} = J^{(k-1)} - \frac{1}{N} g_k^2 \kappa_k + \frac{\tau_A}{\kappa_k} \\
&\text{D-optimality} + \text{OLS}: \; J^{(k)} = J^{(k-1)} - \frac{1}{N} g_k^2 \kappa_k + \tau_D \log\left[\frac{1}{\kappa_k}\right]
\end{aligned} \tag{25}$$

in each forward regression step, where τ_A and τ_D are small positive numbers to regulate the trade-off between model approximation capability and optimal design criteria. Note that the third term in both of (25) is as a result of applying the experimental design criteria of Section 2.2, based on model representation (21). τ_A and τ_D are set by the users and it is shown that model robustness can be improved for a wide range of τ_A or τ_D. Alternatively, τ_A or τ_D can be determined empirically using cross validation from another dataset.

Generally, extra parameters affect modelling performance and it is advantageous to keep the extra tuning parameters to the minimum. An improved model selective criteria is simply the mean squares of LOO errors $E[(\zeta^{(-)}(i))^2]$, which does not include any tuning parameter. The combined LOO errors-based ROLS algorithm has been introduced (Hong, Sharky, and Warwick 2003; Chen, Hong, Harris, and Sharkey 2004), in which

$$J^{(k)} = E\left[(\zeta^{(-)}(i))^2\right] \tag{26}$$

are calculated efficiently. Note that $E[(\zeta^{(-)}(i))^2]$ directly measures the model generalisation capability and has not lost discriminative power in selecting terms, as happens with (10).

In parallel to each of the variants of the orthogonal forward selection algorithm based on A-optimality, D-optimality or LOO errors, the Bayesian regularisation can be efficiently implemented. The regularisation parameters $\mu_i = (\alpha_i/\beta)$ can be optimised using some efficient recursive formula through the evidence maximisation procedure based on (21) and (22), following an evidence maximisation procedure (MacKay 1991; Chen, Hong, and Harris 2003b; Chen et al. 2004).

3.3 *Other stepwise selection algorithms*

Other subset selection algorithms have also been researched. The least squares parameter estimator is also the MLE when the noise is Gaussian distributed with a constant variance. In practice, the Gaussian noise assumption may be violated, e.g. the data samples have outliers. The general method of M-estimation (Huber 1981) is well established in order to tackle outliers in observational data. The M-estimator-based orthogonal selection algorithm has been proposed for robust model identification (Hong and Chen 2005).

The conventional backward elimination approach removes a model regressor one at a time based on the least deterioration in model fit. The computational cost of backward elimination algorithms decreases dramatically when the candidate model size is small, therefore, the backward elimination as a post-processing procedure is computationally affordable, and this can be used to form hybrid approaches to prune a model that is identified via other approaches. The modification on conventional backward elimination approach has been researched by using some hybrid cost functions between the model fit and one of three terms of A-/D-optimality/(parameter 1-norm in basis pursuit) (Hong, Harris, Brown, and Chen 2004) and LOO errors (Hong and Mitchell 2007).

To reduce the computational cost, or to improve the compactness of the models, a few algorithms have been suggested (e.g. Li, Peng, and Irwin 2005; Li, Peng, and Bai 2006). Iterative algorithms have also been proposed where both the forward model selection and backward model refinement are implemented (e.g. Adeney and Korenberg 2000; Li et al. 2006). In Li et al. (2006),

a regression context is defined to enable the forward model selection and backward model refinement within one integrated analytic framework. To further reduce the computational compleity and to improve the model compactness for neural models with tunable parameters, a hybrid algorithm is proposed to simultaneously select model structure and parameter optimisation (Peng, Li, and Huang 2006). Recently, orthogonal forward-selection algorithmic designs have employed an individually tuned diagonal covariance matrix, rather than a fixed common variance (Chen, Hong, Wang, and Harris 2005b), as well as tunable RBF centres instead of restricting RBF centres to input training data points (Chen, Hong, and Harris 2005a). In these approaches (Chen et al. 2005a,b; Peng et al. 2006), the network growing and parameter optimisation are performed within an integrated analytic framework, significantly increasing the modelling capability of a minimal kernel model.

4. Model construction using convex optimisation algorithms

Despite the popularity and great efficiency of the stepwise subset selection algorithms in practical data modelling, the Achilles heel is that the final model is not optimal, and the algorithms can only yield suboptimal solutions. With the advent of increasing affordable computing resources, efforts on less greedy non-linear modelling algorithms have attracted much interest, such as convex optimisation-based algorithms. One of the recent topics in the area of machine learning is the support vector machine (SVM) as a tractable algorithm for classification and regression, based on the (SRM) principle.

4.1 *The structure risk minimisation principle*

In machine learning, a model generalisation measure is used to describe the capacity of a learning machine (Vapnik 1995, 1998; Müller, Mika, Rätsch, Tsuda, and Schölkopf 2001). The SRM principle or Vapnik–Chervonenkis (VC) theory has been introduced for a binary classifier. The optimal model is derived based on the principle of minimising an upper bound of the model's generalisation error (structure risk) given as (28) below. The concept of the model complexity is expressed by the VC dimension ν of the hypothesis space $\mathcal{H}$. Generally, a VC dimension ν is a scalar value, which measures the capability or the expressive power of the hypothesis space $\mathcal{H}$. A set of k hypothesis spaces can be expressed by

$$\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_k \tag{27}$$

with non-decreasing VC dimensions. Suppose that in each hypothesis space $\mathcal{H}_j$, $j=1,\ldots,k$, a model f_k has been found with the minimum empirical risk $R_{emp}(f)$ over the training set. An upper bound of the generalisation error can be expressed by

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{\nu(\ln(2N/\nu)+1) - \ln(\delta/4)}{N}} \tag{28}$$

where $0<\delta<1$ holds with probability of at least $1-\delta$ for $N>\nu$.

The SRM principle chooses the hypothesis space $\mathcal{H}_j$ such that the above upper bound is minimised. Similar to the other model selective criteria of (10), the best model is chosen with a trade-off between empirical errors over estimation dataset and model complexity. The key difference is that model complexity is given as VC dimension ν, which may or may not be related to m, the number of terms in a linear-in-the-parameters model.

The SVM is based on the SRM principle and the theory of the reproducing kernel Hilbert space (RKHS) kernel functions (Aronszajin 1950; Debnath and Mikusinski 1998; Smola and Schlkopf 1998; Gao, Harris, and Gunn 2001; Bartlett 2003). Note that as the VC dimension is difficult to compute it is impractical to implement (28) in practise. Assisted by a 'kernel trick', the SVM approximately minimises an upper bound on the generalisation error (Vapnik 1995, 1998) via minimising the norm of weights in a feature space (Vapnik 1995, 1998). Consider initially a two-class training dataset $\{\mathbf{x}(j), y(j)\}, y(j)\in[-1,1]$, $j=1,\ldots,N$, a hyperplane classifier is represented by $\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x})+b=0$, where $\boldsymbol{\varphi}(\mathbf{x})$ maps the data $\mathbf{x}$ to a feature space $\mathcal{F}$ with dimension $N_{\mathcal{F}}$, which may be unknown and/or even infinite. Thanks to the celebrated Mercer's theorem (Mercer 1909), the difficulty of having to work with the unknown feature space $\mathcal{F}$ can be avoided through the 'kernel trick': for some feature space $\mathcal{F}$ the inner products are calculated without explicitly knowing φ, but via a kernel function[1], i.e. $k(\mathbf{x}(i),\mathbf{x}(j))=\boldsymbol{\varphi}(\mathbf{x}(i))^T\boldsymbol{\varphi}(\mathbf{x}(j))$. A typical example of kernel function is the Gaussian RBF where $k(\mathbf{x}(i),\mathbf{x}(j))=\exp((-\|\mathbf{x}(i)-\mathbf{x}(j)\|^2/\tau^2))$.

Aside from the kernel trick, the invention of SVM is also due to another useful result (Vapnik 1998) of linking the VC dimension with the margin, defined as the minimal distance of a data sample to the decision surface. The margin can be calculated as $(2/\|\mathbf{w}\|)^2$ and an upper bound of the VC dimension is $\mu\leq\|\mathbf{w}\|^2R^2+1$ (Müller et al. 2001), where R is the radius of the smallest ball around the training data, which is fixed for a given dataset. This means that the second term in (28) (the model complexity term) can

be minimised via the minimisation of $\|\mathbf{w}\|^2$. Define a set of slack variables $\boldsymbol{\xi} = [\xi(1), \ldots, \xi(N)]^T$. The SVM can be formulated as a quadratic optimisation problem

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{j=1}^{N}\xi(j) \tag{29}$$

subject to $y(j)(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}(j)) + b) \le 1 - \xi(j)$, $\xi(j) \ge 0, 1 \le j \le N$, where $C > 0$ is the regularisation constant.

The SVM for classification has been comprehensively reviewed (Müller et al. 2001). There is an abundance of more recent publications on SVM, mainly for pattern classification, in all major neural networks and machine learning journals. In the following, we concentrate on the extension of SVM to the regression problem, referred to as the SVR (Smola 1998; Smola and Schlkopf 1998).

4.2 *Support vector regression*

Consider using a model

$$\hat{y} = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b \tag{30}$$

based on the training dataset $\{\mathbf{x}(j), y(j) \in \Re, j = 1, \ldots, N$. Define $\zeta = y - \hat{y}$, a common choice for the loss function in the SVR is an ε-insensitive loss function (Vapnik 1998) given by

$$\rho(\zeta) = \begin{cases} 0 & \text{if } |\zeta| < \varepsilon. \\ |\zeta| - \varepsilon & \text{otherwise.} \end{cases} \tag{31}$$

where $\varepsilon > 0$ is a user-defined threshold for controlling the SV's (support vector) complexity. The use of the ε-insensitive loss function in SVR is equivalent to the formulation of a quadratic optimisation problem (Smola and Schlkopf 1998)

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{j=1}^{N}(\xi(j) + \xi^*(j)) \tag{32}$$

subject to

$$\begin{cases} y(j) - \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}(j)) - b \le \varepsilon + \xi(j), \quad j = 1, \ldots, N, \\ \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}(j)) + b - y(j) \le \varepsilon + \xi^*(j), \quad j = 1, \ldots, N. \\ \xi(j), \xi^*(j) \ge 0, \quad j = 1, \ldots, N \end{cases} \tag{33}$$

where $\boldsymbol{\xi} = [\xi(1), \ldots, \xi(N)]^T$ and $\boldsymbol{\xi}^* = [\xi^*(1), \ldots, \xi^*(N)]^T$ are slack variables.

The constant C determines the trade-off between the complexity of the model and the modelling error over the training set. Using the standard dualisation method, a Lagrangian can be represented by (Smola and Schlkopf 1998):

$$\begin{aligned} &Lgrg(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}, \boldsymbol{\beta}^*) \\ &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{j=1}^{N}(\xi(j) + \xi^*(j)) \\ &\quad - \sum_{j=1}^{N}\alpha(j)(\varepsilon + \xi(j) - y(j) + \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}(j)) + b) \\ &\quad - \sum_{j=1}^{N}\alpha^*(j)(\varepsilon + \xi^*(j) + y(j) - \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}(j)) - b) \\ &\quad - \sum_{j=1}^{N}(\beta(j)\xi(j) + \beta^*(j)\xi^*(j)) \end{aligned} \tag{34}$$

where $\boldsymbol{\alpha} = \{\alpha(j) \ge 0 | j \in (1, \ldots, N)\}$, $\boldsymbol{\alpha}^* = \{\alpha^*(j) \ge 0 | j \in (1, \ldots, N)\}$, $\boldsymbol{\beta} = \{\beta(j) \ge 0 | j \in (1, \ldots, N)\}$, $\boldsymbol{\beta}^* = \{\beta^*(j) \ge 0 | j \in (1, \ldots, N)\}$ are sets of Lagrange multipliers.

It follows from the saddle point condition (Minoux 1986) that the partial derivatives of the Lagrangian given by (34) with respect to $\mathbf{w}$, b, $\xi(j)$ and $\xi^*(j)$, $j = 1, \ldots, N$ have to vanish for optimality, i.e.

$$\frac{\partial Lgrg}{\partial b} = \sum_{j=1}^{N}(\alpha(j) - \alpha^*(j)) = 0, \tag{35}$$

$$\frac{\partial Lgrg}{\partial \mathbf{w}} = \mathbf{w} - \sum_{j=1}^{N}(\alpha(j) - \alpha^*(j))\boldsymbol{\varphi}(\mathbf{x}(j)) = 0, \tag{36}$$

$$\frac{\partial Lgrg}{\partial \xi(j)} = C - \alpha(j) - \beta(j) = 0, \quad j = 1, \ldots, N, \tag{37}$$

$$\frac{\partial Lgrg}{\partial \xi^*(j)} = C - \alpha^*(j) - \beta^*(j) = 0, \quad j = 1, \ldots, N. \tag{38}$$

Substituting (35)–(38) into (34) and making use of the kernel functions

$$k(\mathbf{x}(i), \mathbf{x}(j)) = \boldsymbol{\varphi}(\mathbf{x}(i))^T\boldsymbol{\varphi}(\mathbf{x}(j)), \quad i \in \{1, \ldots, N\}, j \in \{1, \ldots, N\} \tag{39}$$

yields a dual quadratic programming (QP) optimisation problem as (Smola and Schlkopf 1998)

$$\begin{aligned} &\max_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*}\left\{-\frac{1}{2}\sum_{i,j=1}^{N}[\alpha(i) - \alpha^*(i)][\alpha(j) - \alpha^*(j)]k(\mathbf{x}(i), \mathbf{x}(j))\right. \\ &\left. -\varepsilon\sum_{j=1}^{N}(\alpha(j) + \alpha^*(j)) + \sum_{j=1}^{N}(\alpha(j) - \alpha^*(j))y(j)\right\}, \\ &\text{subject to} \begin{cases} 0 \le \alpha(j) \le C, \quad j = 1, \ldots, N, \\ 0 \le \alpha^*(j) \le C, \quad j = 1, \ldots, N, \\ \sum_{j=1}^{N}(\alpha(j) - \alpha^*(j)) = 0. \end{cases} \end{aligned} \tag{40}$$

Note that $\beta(j)$ and $\beta^*(j)$ does not appear in the dual objective (40). By applying (36) to (30), the SVR model (30) can be rewritten as

$$f(\mathbf{x}) = \sum_{j=1}^{N} [\alpha(j) - \alpha^*(j)] k(\mathbf{x}, \mathbf{x}(j)) + b, \tag{41}$$

which is independent of the dimension of the feature space $\mathcal{F}$, but only depends on the number of data samples with non-zero $[\alpha(j) - \alpha^*(j)]$.

The optimal solution has to satisfy the Karush–Kuhn–Tucker (KKT) conditions as (Smola and Schlkopf 1998)

$$\begin{cases} \alpha(i)\left[\varepsilon + \xi(i) - y(i) + \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) + b\right] = 0, \\ \alpha^*(i)\left[\varepsilon + \xi^*(i) + y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) + b\right] = 0. \end{cases} \tag{42}$$

and

$$\begin{cases} (C - \alpha(i))\xi(i) = 0, \\ (C - \alpha^*(i))\xi^*(i) = 0. \end{cases} \tag{43}$$

From (42) and (43), the following conclusions can be drawn:

(1) $\alpha(i)\alpha^*(i) = 0$. This means that $\alpha(i)$ and $\alpha^*(i)$ cannot be non-zero simultaneously. This can be proven by contraction. If $\alpha(i)\alpha^*(i) \neq 0$, then from (42) to yield $2\varepsilon + \xi(i) + \xi^*(i) = 0$ is impossible.

(2) When $|y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) - b| < \varepsilon$, both $\alpha(i)$ and $\alpha^*(i)$ have to be zeros.

(3) When $|y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) - b| \geq \varepsilon$, $\alpha(i)$ or α^*(i) may be non-zeros. Moreover, if $|y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) - b| > \varepsilon$, then from (42), any of $\xi(i)$ and $\xi^*(i)$ has to be non-zero, yielding that the corresponding $\alpha(i)$ or $\alpha^*(i)$ equals to C.

(4) For any data point $\mathbf{x}(i)$, $i \in \{1, \ldots, N\}$ satisfying either $\alpha(i) \in (0, C)$ or $\alpha^*(i) \in (0, C)$, it can be derived from (43) that either $\xi(i) = 0$ or $\xi^*(i) = 0$, respectively.

From (42), it is seen that $|y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) - b| = \varepsilon$.

Applying the fourth point from above, the threshold b can be computed by any of the following equations (Smola and Schlkopf 1998)

$$\begin{cases} b = y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) - \varepsilon & \text{if } \alpha(i) \in (0, C), \\ b = y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) + \varepsilon & \text{if } \alpha^*(i) \in (0, C). \end{cases} \tag{44}$$

From the above points, it is clear that the SVR has the property of sparsity in that only the data points $\mathbf{x}(i)$ with non-zero values of $(\alpha(j) - \alpha^*(j))$ are included in (41). These data points are referred to as the support vectors (SV) and satisfy $|y(i) - \sum_{j=1}^{N} (\alpha(j) - \alpha^*(j)) k(\mathbf{x}(i), \mathbf{x}(j)) - b| \geq \varepsilon$. It is possible to control the number of the SVs by setting a proper value of ε and obtain a final model with a small number of SVs.

The optimisation algorithm of solving the dual (40) has been discussed (Smola and Schlkopf 1998). The training of SVM requires solving a convex QP problem (Vanderbei 1994) and the global optimum is ensured. A problem with standard QP solvers is that they become computationally inefficient for large-sized datasets. One of the efforts of overcoming this problem is the so-called chunking that operates on a working set, a fixed size subset of the training dataset, in order to find the SV for the working set. The process iterates and in each iteration a set of worst input data that violates the optimality condition of the current estimator is chosen as the new working set for the next iteration. For maximal computational efficiency, the sequential minimal optimisation (SMO) has been introduced for classification (Platt 1999) and adopted in SVR (Smola and Schlkopf 1998). The SMO breaks a large QP problem into a series of the smallest possible QP problems of size two, for which the analytical solution is available rather than using the time consuming numerical QP solver as an inner loop. An improved SMO–SVR was proposed to further improve the efficiency of Smola's SMO–SVR (Shevade, Keerthi, Bhattacharyya, and Murthy 2000).

Some possible issues associated with SVR application in system identification have been discussed (Drezet and Harrison 1998). Although the SVR has the properties of sparsity control using ε-insensitive loss function, it was found that the final SVR model may still have a large size despite the use of ε-insensitive loss (Drezet and Harrison 1998; Lee and Billings 2002). There is current research into improving the sparsity of the SVM (Burges 1996; Downs, Gates, and Masters 2001). An empirical study has shown that compared to SVR, the OLS tends to derive much smaller models, yet may be less robust to noise in low signal-to-noise ratio scenarios. The SVM–OLS was proposed to reduce the model size and retain noise rejection of the final model (Lee and Billings 2002). Finally, the LSSVM is an attractive approach for non-linear system identification (Suykens et al. 2002; Espinoza et al. 2005b; Goethals et al. 2005). This is partly because the original LSSVM is *not* solved using a QP solver, but as a much simpler regularised least squares algorithm. However, LSSVM does not lead to a sparse representation; this disadvantage may be alleviated by limiting the search space to a subset of

datasets (Hoegaerts, Suykens, Vandewalle, and De Moor 2005). Alternatively, an hierarchical modelling strategy is adopted to achieve the desirable sparsity representation by using QP (Pelckmans, Suykens, and De Moor 2005).

4.3 *l_1 norm regularisation*

A difference between the orthogonal forward selection (OFS) procedure and the SVR is that OFS algorithms start from an empty model, whereas the SVR algorithm gradually eliminates the data samples from the model. The sparsity of the SVR is achieved by the QP algorithm with the property of global optimality. An alternative QP-based sparse model construction algorithm is the 'least absolute shrinkage and selection operator' (LASSO) algorithm (Tibshirani 1996). The LASSO algorithm uses l_1 norm regularisation and retains some good features from both subset selection and ridge regression (Tibshirani 1996). Recall the regularised parameter estimation objective function $J_R = [\mathbf{y} - \mathbf{\Phi\theta}]^T[\mathbf{y} - \mathbf{\Phi\theta}] + \mu\boldsymbol{\theta}^T\boldsymbol{\theta}$ in Section 2, where the l_2 norm is used to penalise the models with high magnitude of parameter estimator. Consider using an alternative objective function of $J_{R1} = [\mathbf{y} - \mathbf{\Phi\theta}]^T \times [\mathbf{y} - \mathbf{\Phi\theta}] + \mu\|\boldsymbol{\theta}\|_1$, where $\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{m} |\theta_i|$. The l_1 norm regularisation parameter μ is effectively used in tuning the final model size, since more model parameters have exactly zero values as μ is reduced.

Some recent researches examine the penalty function using the more general bridge function as $J_B = [\mathbf{y} - \mathbf{\Phi\theta}]^T[\mathbf{y} - \mathbf{\Phi\theta}] + \mu\|\boldsymbol{\theta}\|_\rho$, especially for $0 < \rho \leq 1$ (Knight and Fu 2000). It has been shown that for $0 < \rho \leq 1$, under appropriate regularity conditions, the limiting distributions can have positive probability mass at 0 when the true value of the parameter is zero. This result provides a theoretic justification for the use of bridge estimators to distinguish between covariates whose coefficients are exactly zero and those are non-zeros. Thus, by appropriate choice of ρ, bridge estimator can combine variable selection and parameter estimation within a single step.

The l_1 norm regularisation from Bayesian point of view is that the parameters have an exponential distribution prior (Tibshirani 1996). The sparse model construction has been researched in the closely-related area of signal approximation (Mallat 1989; Daubechies 1992). In signal approximation, a signal is represented using linear superposition of a minimum number of waveforms, called atoms, that is selected from a large collection of waveforms, called the dictionary. Sparse approximation techniques include the soft-thresholding (Donoho 1995), the matching pursuit (Mallat and Zhang 1993) and the basis pursuit (Chen, Donoho, and Saunders 1998). The LASSO algorithm is equivalent to the basis pursuit (BP) algorithm of signal approximation theory (Chen et al. 1998). If the objective function J_{R1} is modified such that (1) the basis functions are restricted as the Mercer kernel basis functions and (2) the first term in J_{R1}, the approximation error l_2 norm, is replaced by the RKHS norm of the approximation error as induced by the kernel, then there is an exact equivalence between the modified BP and SVR (Girosi 1998). More analysis on general statistical leaning machine including regularisation network theory, sparse approximation and SVM can be found (Evgeniou, Pontil, and Poggio 2000; Hastie, Tibshirani, and Friedman 2002).

Like SVR, the model construction using the objective function of J_{R1} can be formulated as a QP algorithm and becomes computationally inefficient for large-sized datasets. This problem has been attacked successfully by the least angle angle regression (LARS) algorithm, which is closely related to forward regression and LASSO.

5. Input selection algorithms

The identification of non-linear systems involves learning a minimal model representation using finite datasets for *a priori* unknown systems. The structural determination of any model includes the selection of appropriate causal input variables and the candidate set of basis functions. For the actual system output, some input variables may be redundant or would become insignificant if some other input variables were present in the model. The effects of overparametrisation on the system dynamics behaviour have been empirically studied using the qualitative methods and it was found that the overparametrisation in the input lag has the worst effect (Zheng and Billings 1999, 2002). The input selection of choosing the minimum number of relevant inputs as a preprocessing procedure should improve the modelling performance. Optimal input selection is an intractable task as the system input/output relationship is non-linear and the input variables are generally not independent.

Mutual information quantifies the dependence between two random variables and is a theoretically suitable measure of input selection (Battiti 1994). Practical algorithms based on mutual information have been introduced (Zheng and Billings 1996; Kwak 2002; Bowden, Dandy, and Maier 2005). These are greedy algorithms that select input variables one at a time taking into account the mutual information between the candidate input variable,

output variable and that of the set of the selected variables in previous forward stage. The disadvantage of the mutual information-based approach is that some probability density estimator is required and subsequently the requirement on the amount of the data for estimation is high. Alternatively, some simple elimination algorithm has been introduced to detect redundant inputs based on a clustering algorithm and some rules based on the properties of a function (Sindelář and Babuška 2004).

The piecewise local linear model-based input selection approach has been introduced (Mao and Billings 1999). Using the Taylor expansion, (1) can be approximated as

$$y(t) = f(\mathbf{x}(t_0)) + \sum_{k=1}^{n} \frac{\partial f(\mathbf{x}(t))}{\partial x_k(t)}|_{x=x(t_0)}(x_k(t) - x_k(t_0)) + \text{higher order terms} + e(t) \quad (45)$$

where n is the total number of input variables. If the observation vector $\mathbf{x}(t)$ falls into a small neighbourhood of $\mathbf{x}(t_0)$, the input/output relationship is dominated by the first-order derivatives of $f(\bullet)$, then it may be assumed that the *higher order terms* in (45) are negligible and a local linear model would be appropriate.

The algorithm of (Mao and Billings 1999) involves dividing the input space into sub-regions and dataset D_N into different groups according to the sub-regions. For each sub-region, the linear model form is used and the forward OLS algorithm is used to select the subset of input variables. The union of selected input variables are used in the final model construction. A fully automatic piecewise local linear model-based input selection algorithm has been introduced where the piecewise local linear modelling is integrated in each forward stage of the forward OLS algorithm (Hong and Harris 2001b).

An effective approach for simultaneously reducing both the input variables and the basis functions is to base the model on an additive decomposition approach via the well-known analysis of variance (ANOVA) expansion (Bossley 1997)

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_{i,j}(x_i, x_j) + e(\mathbf{x}) \quad (46)$$

where each of f_i and $f_{i,j}$ is linear in the parameters model with the basis functions based on sub-vectors of the input vector. As the orthogonal selection algorithms select the most significant terms to compose a parsimonious model, some of the terms will not be present in the final model and only the significant input variables will be included.

6. On-line system identification algorithms

The model-construction algorithms using orthogonal selection and SVR are off-line algorithms that can access the whole dataset D_N in training. The recursive parameter-estimation algorithms, e.g. the recursive least squares algorithm (RLS) and prediction error algorithm, are directly applicable to the linear in the parameters models (Ljung and Söderström 1983; Söderström and Stoica 1989). A model constructed using off-line algorithms can be applied to another dataset that is similar to the estimation dataset, with the model structure fixed and the parameters updated using linear recursive algorithms. In the case that the new data exhibit characteristics significantly different from the estimation dataset, the model has to be retrained. Alternatively, on-line system-identification algorithms are an important class of model construction algorithms that deal with model-structure and/or parameter updating on the arrival of a new data sample. Suppose that a model is identified using the data samples collected up to time $(t-1)$. Given a new data sample at t and the *a priori* known model of $(t-1)$, the on-line model-identification algorithms address the problem as how to update the model by including the innovation induced by new data sample. On-line system identification algorithms are advantageous in that the model is updated following the arrival of the new data sample rather than to relearn the model from scratch.

An essential characteristic of on-line non-linear system-identification algorithms is the capability of varying model structure to adapt the new data samples. The resource allocating network (RAN) (Platt 1991) is an on-line-identification algorithm to grow Gaussian RBF networks. The growing criteria and parameter adaptation were extended further, in its variant the RAN extended Kalman filter (RANEKF) (Kadirkamanathan and Niranjan 1993). Similarly, the variable neural network has been introduced and applied in adaptive control (Liu, Kadirkamanathan, and Billings 1999). A minimal resource allocating network (MRAN) was proposed that combines RAN with a procedure of pruning the redundant neurons (Yingwei, Sundarajan, and Saratchandran 1998). Recently, the growing and pruning RBF (GAP–RBF) networks have been introduced, which used the concept of 'significance' of a neuron as defined as its contribution, averaged over all the past data, to the model output (Huang, Saratchandran, and Sundaraja 2004a). The GAP–RBF algorithm is designed to achieve a high computational efficiency, which is a critical requirement for on-line algorithms.

The time constraint imposed on any on-line learning scenario makes the task of choosing the

optimal model structure a difficult problem. In practice, in determining the growing and pruning strategy in various on-linear algorithms, various thresholds based on the required modelling accuracy have to be defined by the user, and approximation formulae for fast calculation are often used to speed up the calculation. Although the strategy often works well as demonstrated by empirical studies (Huang et al. 2004a), the connection between the growing and pruning strategy and the model generalisation is not well studied with few exceptions in some algorithms based on the RKHS theory, e.g. the incremental projection learning (Sugiyama and Ogawa 2001) and kernel RLS (Engel, Mannor, and Meir 2004). The kernel RLS-based model is essentially a growing kernel model with a built in on-line sparsification procedure for computational efficiency. The computational complexity for kernel algorithm scales with the sparsity of the kernel machines. The general applicability of kernel algorithms to on-linear applications is critically dependent on the sparsity of the obtained kernel solutions, since this would also help to reduce the computation cost in the on-line sparsification.

7. Applications

Linear-in-the-parameter models have been widely applied for monitoring, controlling and supervisioning across all traditional engineering sectors, like mechanical (Parlitz et al. 2004; Govindhasamy, McLoone, Irwin, French, and Doyle 2005; Altinkok 2006) electrical and electronic (Park et al. 1991; Leva and Piroddi 2002), chemical (Soumelidis and Stobart 2006), energy and power (Glass and Franchek 1999; Jurado 2004; Li, Thompson, and Peng 2004; Basso, Giaue, Groppi, and Zappa 2005), aerospace and aeronautical (Faller and Schreck 1996), civil (Flood and Kartam 1998), and environmental systems (Nunnari et al. 2004; Peng et al. 2004). More recently, applications in newer areas are being reported including communication networks (Chen, Gibson, Cowan, and Grant 1991; Clarkson 1999), biomedical, biochemical and life systems (Ma and Wang 2000; Gamero, Armentano, Barra, Simon, and, Levenson 2001; Sargantanis and Karim 2004; Karayiannis et al. 2006), as well as other sectors other than engineering, such as financial (Marose 1990), social (Garson 1991), health care and medical (Dybowski and Gant 2001). Space prevents a comprehensive survey of such a vast literature survey and so only a few detailed examples are included here.

Among the various linear-in-the-parameter structures available, NARX/NARMAX is one of the earliest and perhaps most widely-used model types, with many successful industrial applications reported. For example, it has been used in the modelling and control of power systems, such as internal combustion engine (Glass and Franchek 1999), automotive diesel engine (Billings, Chen, and Backhouse 1989), power plant gas turbine and micro-turbine modelling (Basso et al. 2005; Jurado 2005), magneto-rheological damping devices (Leva and Piroddi 2002) and fuel cell plants (Jurado 2004), modelling and control of longitudinal vehicle dynamics (Kalkkuhl, Hunt, and Fritze 1999), identification of pre-sliding friction dynamics (Parlitz et al. 2004), modelling of a pH waste water neutralisation process (Luo, Morris, Karim, Martin, and Hong 1996), dynamic modelling of three-way catalysts (Soumelidis and Stobart 2006), and air pollution modelling and control (Peng et al. 2004; Soumelidis and Stobart 2006). It has also been used to model arterial wall dynamics in animals (Gamero et al. 2001), and modelling and control of a batch Bacillus subtilis fermentation process (Sargantanis and Karim 2004).

The generalised single hidden-layer neural network, an important type of linear-in-the-parameter models, covers a number of well-known neural network paradigms. These include the most popular ones like the RBF neural networks, Volterra neural networks and B-spline neural networks. These have proved powerful modelling tools, with significant impact reported in the literature on signal processing and pattern recognition (Clarkson 1999; Xie and Leung 2005; Lin, Chang, and Lai 2007), time-series prediction (Leung, Lo, and Wang 2001), and non-linear system modelling and control (Sanner and Slotine 1992; Irwin, Warwick, and Hunt 1995; Lewis and Parisini 1998; Ge, Hang, Lee and Zhang 2001; Liu 2001; Huang, Tan, and Lee 2006) with many successful applications in various engineering disciplines (Faller and Schreck 1996; Vemuri, Polycarpou, and Diakourtis 1998; Flood and Kartam 1998; Wilson, Irwin, and Lightbody 1999; Soumelidis and Stobart 2006) as well as social, medical and other sectors (Marose 1990; Harrison and Garson 1991; Ma and Wang 2000; Dybowski and Gant 2001; Kennedy 2005).

For example, Govindhasamy et al. (2005) reports collaborative research with Seagate Technology Media (Ireland) Ltd., the world's largest manufacturer of hard-disc drives, with 159.2 million units shipped in the 12 months ended in June 2007. The aim was optimisation of a grinding process used to machine the aluminium substrate disks, the main component of a disk drive, to a desired thickness in order to minimise the number of out-of-specification disks produced. The process involved 12 parts being simultaneously ground at each machine cycle. A proprietary thickness control algorithm, employing thickness measurements before and after grinding, was used to calculate the average stock removal rate for each machine cycle.

Unfortunately, this did not adequately account for the non-linear variation of the cycle-to-cycle removal-rate, and non-linear system identification was used to enable removal-rate prediction. A single hidden-layer generalised MLP, with its additional direct connections from the inputs to the output, proved highly effective in capturing both the linear and the non-linear dynamic components of removal rate in tests on practical grindstone data. Further, using the resultant neural NARX model for thickness control yielded much tighter process control than the existing proprietary control scheme. In this application, the generalised MLP provided an ideal modelling tool, as it can be initialised as a linear model and then adapted to produce the required non-linear representation.

Along with the growing popularity of SVM, successful applications have been reported in engineering system modelling and control (Drezet and Harrison 1998; Iplikci 2006), such as modelling of proton exchange membrane fuel cell (Zhong, Zhu, and Cao 2006) and quality monitoring of a plastic injection moulding process (Ribeiro 2005), etc. It has also been widely used in bioinformatics (Bradford and Westhead 2005; Tothill et al. 2005), Geo- and environmental sciences (Kanevski et al. 2004; Khan and Coulibaly 2006), and many other areas (Hong 2006), though one of the major recent contributions is on bioinformatics for classification of high throughput 'nomics' datasets, e.g. Mitra and Hayashi (2006); Xia and Li (2007).

Finally, some other linear-in-the-parameters (grey-box) models have also been explored, where the model structure, non-linear model terms or network activation functions are mostly system specific (Lorito 1999; Li et al. 2004; Raghavan et al. 2005; Bohlin 2006; Bohlin 2006; Li and Peng 2006). For example, non-linear system identification was successfully applied to the environmentally important topic of NO_x emissions modelling and prediction of a 200-MW coal-fired power generation plant (Li et al. 2004). The fundamental mechanisms governing NO_x formation combined with system identification ideas produced a semi-physical, grey-box regression model. While both linear and non-linear ARX models gave comparable one step ahead, short-term prediction performances, only the grey-box model was capable of open-loop predictions of NO_x emissions spanning several weeks. It has also been reported that the inherent symmetry properties of the system can be incorporated into the linear-in-the-parameter models to improve the performance (Aguirre, Lopes, Amaral, and Letellier 2004; Espinoza, Suykens, and De Moor 2005a; Chen, Wolfgang, Harris, and Hanzon 2007), system eigenvalues can be used to choose the model types (Aguirre, Coelho, and Correa 2005), and simple system *a priori* information like steady-state relations of variables has been used in the identification of non-linear models for a Buck Converter (Aguirre, Donoso-Gauia, and Santos-Filho 2000). System specific neural network models have also been studied (Li 2005; Li and Peng 2006) with application to power plants and chemical processes, and the activation functions in the network are system specific, aiming to improve the model interpretability and generalisation performance (Connally et al. 2005, 2007).

While the linear-in-the-parameter models have been widely accepted in industry, it is also worthwhile to note that some other model types have also become popular alternatives, for example, the local models and local model networks for modelling and control of complex processes and systems with different operating conditions (Murray-Smith 1994; Gray, Murray-Smith, Li, and Sharman 1996; McGinnity and Irwin 1996; Brown, Irwin, and Lightbody 1997; Townsend and Irwin 2001; Brown, Flynn, and Irwin 2002).

8. Conclusions

In this review article we have covered many of the major advances in linear-in-the-parameter non-linear system-identification algorithms that only utilise data to construct the most parsimonious model of the underlying process. As these algorithms are accepted by industry, the number of industrial applications will increase dramatically as the demand for increasing accuracy and representation of complex dynamic processes continues. Whilst the research fields of kernel methods and associated algorithms such as SVMs are relatively mature for stationary, stochastic processes for which batch data are readily available, on-line, recursive algorithms that simultaneously find model structure and parameterisation are relatively new and present an open field for new and demanding research. Equally, the field of the identification of unknown non-stationary or time-varying processes is both vitally important, but not completely open due to its current intractability. We hope to report progress in both fields in the future!

Acknowledgements

The authors thank the referees and the editors for their constructive comments to improve the quality of the article.

Note

1. Functions that satisfy Mercer's condition, i.e. for any $g(\mathbf{x})$ such that $\int g(\mathbf{x})^2 \mathrm{d}\mathbf{x}$ is finite, then $\int k(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) \times \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}\} \geq 0$.

References

Adeney, K.M., and Korenberg, M.J. (2000), ''Iterative Fast Orthogonal Search Algorithm for MDL-based Training of Generalised Single-layer Network,'' *Neural Networks*, 13, 787–799.

Aguirre, L.A., Coelho, M.C.S., and Correa, M.V (2005), ''On the Interpretation and Practice of Dynamical Differences Between Hammerstein and Wiener Models,'' *IEE Proc. and Control Theory Applications*, 152(4), 349–356.

Aguirre, L.A., Donoso-Garcia, P.F., and Santos-Filho, R. (2000), ''Use of a Priori Information in the Identification of Global Nonlinear Models-a Case Study using a Buck Converter,'' *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, 47(7), 1081–1085.

Aguirre, L.A., Lopes, R.A.M., Amaral, G.F.V., and Letellier, C. (2004), ''Constraining the topology of Neural Networks to Ensure Dynamics with Symmetry Properties,'' *Physical Review E*, 69, 026701–1–026701–11.

Akaike, H. (1974), ''A New Look at the Statistical Model Identification,'' *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Allen, D.M. (1974), ''The Relationship Between Variable Selection and Data Augumentation and a Method for Prediction,'' *Technometrics*, 16, 125–127.

Altinkok, N. (2006), ''Use of Artificial Neural Network for Prediction of Mechanical Properties of α-Al_2O_3 Particulate-reinforced alsi10mg Alloy Composites Prepared by using Stir Casting Process,'' *Journal of Composite Materials*, 40(9), 779–796.

Aronszajin, N. (1950), ''Theory of Reproducing Kernels,'' *Transactions on American Mathematical Society*, 68, 337–404.

Aström, K.J., and Eyhhoff, P. (1971), ''System Identification – a Survey,'' *Automatica*, 7, 123–162.

Aström, K.J., and Wittenmark, B. (1989), *Adaptive Control*, MS: Addison Wesley.

Atkinson, A.C., and Donev, A.N. (1992), *Optimum Experimental Designs*, Oxford: Clarendon Press.

Bai, E.W. (1998), ''An Optimal Two-stage Identification Algorithm for Hammerstein-wiener Nonlinear Systems,'' *Automatica*, 34, 333–338.

Bai, E.W. (2004), ''Decoupling the Linear and Nonlinear Parts in Hammerstein Model Identification,'' *Automatica*, 40, 671–676.

Barron, A.R. (1984), ''Predicted Square Error. A Criterion for Automatic Model Selection,'' in *SelfOrganizing Methods in Modeling*, ed. S. Farlow, Cambridge, MA: MIT Press.

Bartlett, P.L. (2003), ''Reproducing Kernel Hilbert Spaces,'' *Lecture Notes, CS281B/Stat241B, Statistical Learning Theory*, Berkeley, USA: The Computer Science Division, University of California.

Basso, M., Giarre, L., Groppi, S., and Zappa, G. (2005), ''NaRX Models of an Industrial Power Plant Gas Turbine,'' *IEEE Transactions on Control Systems Technology*, 13(4), 599–604.

Battiti, R. (1994), ''Using Mutual Information for Selecting Features in Supervised Neural Net Learning,'' *IEEE Transactions on Neural Networks*, 5(4), 537–550.

Billings, S.A., and Wei, H.L. (2005), ''The Wavelet-narmax Representation: A Hybrid Model Structure Combining Polynomial Models with Multiresolution Wavelet Decompositions,'' *International Journal of Systems Science*, 36(3), 137–152.

Billings, S.A., and Voon, W.S.F. (1987), ''Piecewise Linear Identification of Nonlinear Systems,'' *International Journal of Control*, 46, 215–235.

Billings, S.A., Chen, S., and Backhouse, R.J. (1989), ''The Identification of Linear and Nonlinear Models of a Turbocharged Automative Diesel Engine,'' *Mechanical Systems and Signal Processings*, 3(2), 123–142.

Bohlin, T. (1971), ''On the Problem of Ambiguities in Maximum Likelihood Identification,'' *Automatica*, 7, 199–200.

Bohlin, T. (2006), *Practical Grey-box Process Identification: Theory and Applications*, New York: Springer.

Bossley, K.M. (1997), *Neurofuzzy Modelling Approaches in System Identification*, Ph.D. thesis, University of Southampton, Dept. of ECS.

Bowden, G.J., Dandy, G.C., and Maier, H.R. (2005), ''Input Determination for Neural Network Models in Water Resources Applications. Part 1-background and Methodology,'' *Journal of Hydrology*, 301, 75–92.

Box, G.E.P., and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day Inc.

Bradford, J.R., and Westhead, D.R. (2005), ''Improved Prediction of Protein-protein Binding Sites using a Support Vector Machines Approach,'' *Bioinformatics*, 21(8), 1487–1494.

Broomhead, D.S., and Lowe, D. (1988), ''Multivariable Functional Interpolation and Adaptive Networks,'' *Complex Systems*, 2, 321–355.

Brown, M., and Harris, C.J. (1994), *Neurofuzzy Adaptive Modelling and Control*, Hemel Hempstead: Prentice Hall.

Brown, M.D., Flynn, D., and Irwin, G.W. (2002), ''Multiple Model Nonlinear Control of Synchronous Generators,'' *Transactions on Inst MC, Special Issue on Electrical Power Generation*, 24(3), 215–230.

Brown, M.D., Irwin, G.W., and Lightbody, G. (1997), ''Nonlinear Internal Model Control using Local Model Networks with Application to a pH Process,'' *Proc. IFAC nt. Symp. on Advanced Control of Chemical Processes, AdChem 97*, Canada: Banff, pp. 13–18.

Burges, C.J. (1996), ''Simplified Support Vector Decision Rules,'' *Proc. International Conference on Machine Learning*, pp. 71–77.

Chen, S. (2002), ''Locally Regularised Orthogonal Least Squares Algorithm for the Construction of Sparse Kernel Regression Models,'' *Proceedings of 6th Int. Cof. Signal Processing*, China: Beijing, pp. 1229–1232.

Chen, S., Billings, S.A., and Luo, W. (1989), ''Orthogonal Least Squares Methods and their Applications to Non-linear System Identification,'' *International Journal of Control*, 50, 1873–1896.

Chen, S.S., Donoho, D.L., and Saunders, M.A. (1998), ''Atomic Decomposition by Basis Pursuit,'' *SIAM Journal on Scientific Computing*, 20(1), 33–61.

Chen, S., Hong, X., and Harris, C.J. (2003b), ''Sparse Multioutput Radial Basis Function Network Construction

using Combined Locally Regularised Orthogonal Least Square and D-optimality Experimental Design,'' *IEE Proceedings Control Theory and Applications*, 150(2), 139–146.

Chen, S., Hong, X., and Harris, C.J. (2003a), ''Sparse Kernel Regression Modelling using Combined Locally Regularised Orthogonal Least Squares and D-optimality Experimental Design,'' *IEEE Transactions on Automatic Control*, 48(6), 1029–1036.

Chen, S., Hong, X., and Harris, C.J. (2005a), ''Orthogonal forward Selection for Constructing the Radial Basis Function Network with Tunable Nodes,'' *Proceedings of 2005 Int. Conf. Intelligent Computing*, Hefei, China, pp. 777–786.

Chen, S., Wu, Y., and Luk, B.L. (1999), ''Combined Genetic Algorithm Optimization and Regularized Orthogonal Least Squares Learning for Radial Basis Function Networks,'' *IEEE Transactions on Neural Networks*, 10, 1239–1243.

Chen, S., Gibson, G.J., Cowan, C.F., and Grant, P.M. (1991), ''Reconstruction of Binary Signals using an Adaptive Radial-basis-function Equalizer,'' *Signal Processing*, 20(1), 77–93.

Chen, S., Hong, X., Harris, C.J., and Sharkey, P.M. (2004), ''Sparse Modelling using Orthogonal forward Regression with PRESS Statistic and Regularization,'' *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 34(2), 898–911.

Chen, S., Hong, X., Wang, X.X., and Harris, C.J. (2005b), ''Identification of Nonlinear Systems using Generalised Kernel Models,'' *IEEE Transactions on Control Systems Technology*, 13(3), 401–411.

Chen, S., Wolfgang, A., Harris, C.J., and Hanzo, L. (2007), ''Symmetric Kernel Detector for Multiple-antenna Aided Beamforming Systems,'' *2007 International Joint Conference on Neural Networks (IJCNN)*.

Chiras, N., Evans, C., and Rees, D. (2001), ''Nonlinear Gas Turbine Modelling using Narmax Structures,'' *IEEE Transactions on Instrumentation and Measurement*, 50(4), 893–898.

Clarkson, T. (1999), ''Applications of Neural Networks in Telecommunications,'' *Proceedings of the ERUDIT Workshop on Application of Computational Intelligence Techniques in Telecommunication*, London: Imperial College.

Cohn, D.A., Ghahramani, Z., and Jordan, M.I. (1996), ''Active Learning with Statistical Models,'' *Journal of Artificial Intelligence Research*, 4, 129–145.

Connally, P., Li, K., and Irwin, G.W. (2005), ''Two Applications of Eng-genes Bases Nonlinear Identification,'' in *Proceedings of the 16th IFAC World Congress*, Prague.

Connally, P., Li, K., and Irwin, G.W. (2007), ''Integrated Structure Selection and Parameter Optimisation for Eng-genes Neural Models,'' *Neurocomputing*. (in press) doi:10.1016/j.neucom.2007.06.005.

Cybenko, G. (1989), ''Approximations by Superpositions of Sigmoidal Functions,'' *Mathematics of Control, Signals, and Systems*, 2, 303–314.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia, PA: SIAM.

Debnath, L., and Mikusinski, P. (1998), *Introduction to Hilbert Spaces with Applications*, San Diego, CA: Academic Press.

Donoho, D.L. (1995), ''De-noising by Soft-thresholding,'' *IEEE Transactions on Information Theory*, 41(3), 613–627.

Downs, T., Gates, K., and Masters, A. (2001), ''Exact Simplification of Support Vector Solutions,'' *Journal of Machine Learning Research*, 2, 293–297.

Drezet, P.M.L., and Harrison, R.F. (1998), ''Support Vector Machines for System Identification,'' *Proceedings of UKACC International Conference on Control*, Vol. 1, UK: Swansea, pp. 688–692.

Dybowski, R., and Gant, V. (2001), *Clinical Applications of Artificial Neural Networks*, Cambridge: Cambridge University Press.

Engel, Y., Mannor, S., and Meir, R. (2004), ''The Kernel Recursive Least Squares Algorithm,'' *IEEE Transactions on Signal Processing*, 52(8), 2275– 2285.

Espinoza, M., Suykens, J.A.K., and De Moor, B. (2005a), ''Imposing Symmetry in Least Squares Support Vector Machines Regression,'' *Proc. Joint 44th IEEE Conf. Decision and Control and European Control Conf.* 2005, Seville, Spain, pp. 5716–5721.

Espinoza, M., Suykens, J.A.K., and De Moor, B. (2005b), ''Kernel Based Partially Linear Models and Nonlinear Identification,'' *IEEE Transactions on Automatic Control*, 50(10), 1602–1606.

Evgeniou, T., Pontil, M., and Poggio, T. (2000), ''Regularization Networks and Support Vector Machine,'' *Advances in Computational Mathematics*, 13(1), 1–50.

Eykhoff, P. (1974), *System Identification – Parameter and State Estimation*, Chichester: J. Wiley.

Fabri, S.G., and Kadirkamanathan, V. (2001), *Functional Adaptive Control: An Intelligent Systems Approach*, London, Berlin, Heidelberg: Springer.

Faller, W.E., and Schreck, S.J. (1996), ''Neural Networks: Applications and Opportunities in Aeronautics,'' *Progress in Aerospace Sciences*, 32(5), 433–456.

Flood, I., and Kartam, N. (1998), ''Artificial Neural Networks for Civil Engineers: Advanced Features and Applications,'' *Handbook of Intelligent Control*, Reston, VA: American Society of Civil Engineers.

Friedman, J.H. (1991), '' Multivariate Adaptive Regression Splines,'' *The Annals of Statistics*, 19(1), 1–141.

Friedman, J.H., and Stuetzle, W. (1981), ''Projection Pursuit Regression,'' *Journal of the American Statistical Association*, 76(376), 817–823.

Funahashi, K. (1989), ''On the Approximate Realization of Continuous Mapping by Neural Networks,'' *Neural Networks*, 2, 183–192.

Gamero, L.G., Armentano, R.L., Barra, J.G., Simon, A., and Levenson, J. (2001), ''Identification of Arterial Wall Dynamics in Conscious Dogs,'' *Experimental Physiology*, 86, 519–528.

Gao, Y., and Er, M.J. (2003), ''Online Adaptive Fuzzy Neural Identification and Control of a Class of Mimo Nonlinear Systems,'' *IEEE Transactions on Fuzzy Systems*, 11(4), 462–477.

Gao, J.B., Harris, C.J., and Gunn, S.R. (2001), "On a Class of Support Vector Kernels Based on Frames in Function Hilbert Spaces," *Neural Computations*, 13, 1975–1994.

Garson, G.D. (1991), "A Comparison of Neural Network and Expert Systems Algorithms with Common Multivariate Procedures for analysis of Social Science Data," *Social Science Computer Review*, 9(3), 399–434.

Ge, S.S., Hang, C.C., Lee, T.H., and Zhang, T. (2001), *Stable Adaptive Neural Network Control*, Boston, MA: Kluwer Academic.

Geman, S., Bienenstock, E., and Doursat, R. (1992), "Neural Networks and the Bias/VAriance Dilemma," *Neural Computation*, 4, 1–58.

Gevers, M. (2005), "Identification for Control: From the Early Achievements to the Revival of Experiment Design," *European Journal of Control*, 11, 1–18.

Girosi, F. (1998), "An Equivalence Between Sparse Approximation and Support Vector Machines," *Neural Computation*, 10(6), 1455–1480.

Girosi, F., and Poggio, T. (1990), "Networks and the Best Approximation Property," *Biological Cybernetics*, 63, 169–176.

Glass, J.W., and Franchek, M.A. (1999), "NaRMAX Modelling and Robust Control of Internal Combustion Engines," *International Journal of Control*, 72(4), 289–304.

Goethals, I., Pelckmans, K., Suykens, J.A.K., and De Moor, B. (2005), "Identification of MIMO Hammerstein Models using Least Squares Support Vector Machines," *Automatica*, 41, 1263–1272.

Goethals, I., Van Gestel, T., Suykens, J., Van Dooren, P., and De Moor, B. (2003), "Identification of Positive Real Models in Subspace Identification by using Regularization," *IEEE Transactions on Automatic Control*, 48(10), 1843–1847.

Goodwin, G.C., and Sin, K.S. (1984), *Adaptive Filtering Prediction and Control*, Englewood Cliffs, NJ: Prentice Hall.

Govindhasamy, J.J., McLoone, S.F., Irwin, G.W., French, J.J., and Doyle, R. (2005), "Neural Modelling, Control and Optimisation of an industrial Grinding Process," *Control Engineering Practice*, 13(10), 1243–1258.

Gray, G.J., Murray-Smith, D.J., Li, Y., and Sharman, K.C. (1996), "Nonlinear Model Structure Identification using Genetic Programming," in *Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University July 28–31, 1996*, ed. John R. Koza, CA, USA: Stanford Bookstore, Stanford University, pp. 32–37.

Greblicki, W. (1989), "Nonparametric Orthogonal Series Identification of Hammerstein Systems," *International Journal of Systems Science*, 20, 2355–2367.

Green, P.J., and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.

Harris, C.J., Hong, X., and Gan, Q. (2002), *Adaptive Modelling, Estimation and Fusion from Data: A Neurofuzzy Approach*, Heidelberg: Springer-Verlag.

Harrison, R.F., and Kennedy, R.L. (2005), "Artificial Neural Network Models for Prediction of Acute Coronary Syndromes using Clinical Data from the Time of Presentation," *Annals of Emergency Medicine*, 46, 431–439.

Hastie, T.J., and Tibshirani, R.J. (1996), *Generalised Additive Models*, London: Chapman and Hall.

Hastie, T.J., Tibshirani, R.J., and Friedman, J. (2002), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Hoegaerts, L., Suykens, J.A.K., Vandewalle, J., and De Moor, B. (2005), "Subset Based Least Squares Subspace Regression in Rkhs," *Neurocomputing*, 63, 293–323.

Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12(1), 55–67.

Hong, W.C. (2006), "The Application of Support Vector Machines to forecast tourist Arrivals in Barbados: An Empirical Study," *International Journal of Management*, 23(2), 375–385.

Hong, X., and Chen, S. (2005), "M-estimator and D-Optimality Model Construction using Orthogonal Forward Regression," *IEEE Transactions on System, Man, and Cybernetics, Part B*, 35(1), 155–162.

Hong, X., and Harris, C.J. (2001a), "Neurofuzzy Design and Model Construction of Nonlinear Dynamical Processes from Data," *IEE Proceedings Control Theory and Applications*, 148(6), 530–538.

Hong, X., and Harris, C.J. (2001b), "Variable Selection Algorithm for the Construction of Mimo Operating Point Dependent Neurofuzzy Network," *IEEE Transactions on Fuzzy Systems*, 9(1), 88–101.

Hong, X., and Harris, C.J. (2003), "Experimental Design and Model Construction Algorithms for Radial Basis Function Networks," *International Journal of Systems Science*, 34(14–15), 733–745.

Hong, X., and Mitchell, R.J. (2007), "Backward Elimination Model Construction for Regression and Classification using Leave one Out Criteria," *International Journal of Systems Science*, 38(2), 101–113.

Hong, X., Harris, C.J., Brown, M., and Chen, S. (2004), "Backward Elimination Methods for Associative Memory Network Pruning," *International Journal of Hybrid Intelligent Systems*, 1(1–2), 90–98.

Hong, X., Sharkey, P.M., and Warwick, K. (2003), "Automatic Nonlinear Predictive Model Construction using forward Regression and the PRESS Statistic," *IEE Proceedings and Control Theory Applications*, 150(3), 245–254.

Hornik, K., Stinchcombe, M., and White, H. (1989), "Multilayer Feedforward Networks are Universal Approximator," *Neural Networks*, 2, 359–366.

Huang, G.B., Saratchandran, P., and Sundararaja, N. (2004a), "An Efficient Sequential Learning Algorithm for Growing and Pruning RBF (GAP-RBF)networks," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 34(6), 2284–2292.

Huang, G.B., Zhu, Q.Y., and Siew, C.K. (2004b), "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks," *Proceedings of*

International Joint Conference on Neural Networks (IJCNN2004), Vol. 2, Hungary: Budapest, pp. 985–990.

Huang, S.N., Tan, K.K., and Lee, T.H. (2006), ''Nonlinear Adaptive Control of interconnected Systems using Neural Networks,'' *IEEE Transactions on Neural Network*, 17(1), 243–246.

Huber, P.J. (1981), *Robust Statistics*, New Jersey: J. Wiley.

Iplikci, S. (2006), ''Support Vector Machines-based Generalized Predictive Control,'' *International Journal of Robust and Non-Linear Control*, 16, 834–862.

Irwin, G.W., Warwick, K., and Hunt, K.J. (1995), *Neural Network Applications in Control*, London: IEE.

James, W., and Stein, C. (1961), ''Estimation with Quadratic Loss,'' *Proceedings Fourth Berkeley Symp. Math. Statist. Prob.*, 1, 311–319.

Juditsky, A., Hjalmarsson, H., Benveniste, A., Delyon, B., Ljung, L., Sjöberg, J., and Zhang, Q. (1995), ''Nonlinear Black-box Modelling in System Identification: Mathematical Foundations,'' *Automatica*, 31(12), 1725–1752.

Jurado, F. (2004), ''Experience with Non-linear Model Identification for Fuel Cell Plants,'' *Fuel Cells*, 5, 105–114.

Jurado, F. (2005), ''Non-linear Modeling of Micro-turbines using NARX Structures on the Distribution Feeder,'' *Energy Conversion and Management*, 46(3), 385–401.

Kadirkamanathan, V., and Niranjan, M. (1993), ''A Function Estimation Approach to Sequential Learning with Neural Networks,'' *Neural Computation*, 5(6), 954–975.

Kalkkuhl, J., Hunt, K.J., and Fritz, H. (1999), ''Fem-based Neural-network Approach to Nonlinear Modeling with Application to Longitudinal Vehicle Dynamics Control,'' *IEEE Transactions on Neural Networks*, 10(4), 885–897.

Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Yatsalo, B., and Canu, S. (2004), ''Environmental Data Mining and Modelling Based on Machine Learning Algorithms and Geostatistics,'' *Journal of Environmental Modelling and Software*, 19, 845–855.

Karayiannis, N.B., Mukherjee, A., Glover, J.R., Ktonas, P.Y., Frost, Jr, J.D., Hrachovy, R.A., and Mizrahi, E.M. (2006), ''Detection of Pseudosinusoidal Epileptic Seizure Segments in the Neonatal Eeg by Cascading a Rule-based Algorithm with a Neural Network,'' *IEEE Transactions on Biomedical Engineering*, 53(4), 633–641.

Kavli, T. (1993), ''AsMOD – an Algorithm for Adaptive Spline Modelling of Observation Data,'' *International Journal of Control*, 58(4), 947–967.

Khan, M.S., and Coulibaly, P. (2006), ''Application of Support Vector Machine in Lake Water Level Prediction,'' *Journal of Hydrologic Engineering*, 11(3), 199–205.

Knight, K., and Fu, W.J. (2000), ''Asymptotics for Lasso-type Estimators,'' *Annals of Statistitics*, 28, 1356–1378.

Korenberg, M.J. (1988), ''Identifying Nonlinear Difference Equation and Functional Expansion Representations: The Fast Orthogonal Algorithm,'' *Annals of Biomedical Engineering*, 16, 123–142.

Kwak, N. (2002), ''Input Feature Selection by Mutual Information Based on Parzen Window,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12), 1667–1671.

Lee, K.L., and Billings, S.A. (2002), ''Time Series Prediction using Support Vector Machines, the Orthogonal, and the Regularised Orthogonal Least Squares Algorithms,'' *International Journal of Systems Science*, 33, 811–821.

Leontaritis, I.J., and Billings, S.A. (1985), ''Input-output Parametric Models for Nonlinear Systems – Part 1: Deterministic Nonlinear Systems; Part 2: Stochastic Nonlinear Systems,'' *International Journal of Control*, 41(1), 303–344.

Leung, H., Lo, T., and Wang, S. (2001), ''Prediction of Noisy Chaotic Time Series using an Optimal Radial Basis Function Neural Network,'' *IEEE Transactions on Neural Networks*, 12(5), 1163–1172.

Leva, A., and Piroddi, L. (2002), ''Narx-based Technique for the Modelling of Magneto-rheological Damping Devices,'' *Smart Materials Structures*, 11, 79–88.

Lewis, F.L., and Parisini, T. (1998), ''Guest Editorial: Neural Network Feedback with Guaranteed Stability,'' *International Journal of Control*, 70(3), 337–339.

Li, K. (2005), ''Eng-genes: A New Genetic Modelling Approach for Nonlinear Dynamic Systems,'' in *Proceedings of the 16th IFAC World Congress*, Prague.

Li, K., and Peng, J.X. (2006), ''System Oriented Neural Networks – Problem formulation, Methodology and Application,'' *International Journal of Pattern Recognition and Artificial Intelligence*, 20(2), 143–158.

Li, K., Peng, J., and Bai, E.W. (2006), ''A Two-stage Algorithm for Identification of Nonlinear Dynamic Systems,'' *Automatica*, 42(7), 1189–1197.

Li, K., Peng, J.X., and Irwin, G.W. (2005), ''A Fast Nonlinear Model Identification Method,'' *IEEE Transactions on Automatic Control*, 50(8), 1211–1216.

Li, K., Thompson, S., and Peng, J. (2004), ''Modelling and Prediction of NOx Emission in a Coal-fired Power Generation Plant,'' *Control Engineering Practice*, 12, 707–723.

Lin, B., Lin, B., Chong, F., and Lai, F. (2007), *Higher-order-Statistics-based Radial basis Function Networks for Signal Enhancement*, 18(3), 823–832.

Liu, G.P. (2001), *Nonlinear Identification and Control-A Neural Network Approach*, London: Springer.

Liu, G.P., Kadirkamanathan, V., and Billings, S.A. (1999), ''Variable Neural Networks for Adaptive Control of Nonliear Systems,'' *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 29(1), 34–43.

Ljung, L. (1987), *System Identification: Theory for the User*, New Jersey: Prentice Hall.

Ljung, L., and Vicino, A. (2005), ''Special Issue on Identification,'' *IEEE Transactions on Automatic Control*, 50(10), 1477–1634.

Ljung, L., and Söderström, T. (1983), *Theory and Practice of Recersive Identification*, Cambridge, MA: The MIT Press.

Lorito, F. (1999), ''Identification of a Grey-box Model of Nonlinear Current Transformers for Simulation Purposes,'' *Control Engineering Practice*, 6(11), 1331–1339.

Luh, G.C., and Cheng, W.C. (2004), ''Identification of Immune Models for Fault Detection,'' *Proceedings Instn Mech. Engrs Part I:J Systems and Control Engineering*, 218, 353–367.

Luo, W., Morris, A.J., Karim, M.N., Martin, E.B., and Hong, T. (1996), ''Online Identification of a Ph Waste Water Neutralisation Process using Time-varying Nonlinear Arx Models,'' in *IEEE International Conference on Systems, Man, and Cybernetics, 1996*, Vol. 1. 14–17 Oct., pp. 107–112.

Ma, Q., and Wang, J. (2000), ''Application of Bayesian Neural Networks to Protein Sequence Classification,'' in *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 305–309.

MacKay, D.J.C. (1991), *Bayesian Methods for Adaptive Models. PhD thesis*, USA: California Institute of Technology.

Mackay, D.J. (1997), ''Gaussian Processes: A Replacement for Supervised Neural Networks,'' *Lecture notes at NIPS 1997*, http://www.ra.phy.cam.ac.uk/mackay

Mallat, S.G. (1989), ''A theory for Multiresolutional Signal Decomposition: The Wavelet Representation,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693.

Mallat, S.G., and Zhang, Z. (1993), ''Matching Pursuits with Time-frequency Dictionaries,'' *IEEE Transactions on Signal Processing*, 41(12), 3397–3415.

Mandic, D.P., and Chambers, J.A. (2001), *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures, and Stability*, Chichester: J Wiley.

Mao, K.Z., and Billings, S.A. (1999), ''Variable Selection in Nonlinear Systems Modelling,'' *Mechanical Systems and Signal Processing*, 13(2), 351–366.

Mao, K.Z., and Billings, S.A. (2000), ''Multi-directional Model Validity Tests for Nonlinear System Identification,'' *International Journal of Control*, 73, 132–143.

Mao, K.Z., Billings, S.A., and Zhu, Q.M. (1999), ''A Regularised Least Squares Algorithm for Nonlinear Rational Model Identification,'' *International Journal of Systems Science*, 30, 455–465.

Markovsky, I., Willems, J.C., Rapisarda, P., and de Moor, B.L.M. (2005), ''Algorithms for Deterministic Balanced Subspace Identification,'' *Automatica*, 41, 755–766.

Marose, R.A. (1990), ''A Financial Neural-network Application,'' *AI Expert*, 5(5), 50–53.

Marquardt, D.W. (1970), ''Generalised inverse, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation,'' *Technometrics*, 12(3), 591–612.

McGinnity, S., and Irwin, G.W. (1996), ''Nonlinear State Estimation using Fuzzy Local Linear Models,'' *International Journal of Systems Science*, 28(7), 643–656.

Mercer, J. (1909), ''Functions of Positive and Negative Type and their Connection with the Theory of Integral Equations,'' *Philos. Transactions on Roy. Soc. London*, A 209, 415–446.

Miller, A.J. (1990), *Subset Selection in Regression*, Boca Raton, Florida: Chapman and Hall/CRC.

Minoux, M. (1986), *Mathematical Programming: Theory and Algorithms*, New York: Wiley.

Mitra, S., and Hayashi, Y. (2006), ''Bioinformatics with Soft Computing,'' *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 36(5), 616–635.

Moody, J. (1994), ''Prediction Risk and Architecture Selection for Neural Networks,'' in *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, ed. V. Cherkassky, Berlin, New York: Springer-Verlag.

Müller, K.R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001), ''An introduction to Kernel Based Learning Algorithms,'' *IEEE Transactions on Neural Networks*, 12(2), 181–201.

Murray-Smith, R. (1994), ''A Local Model Network Approach to Nonlinear Modelling,'' PhD thesis., *Dept. of Computer Science*, Glasgow: University of Strathclyde.

Murray-Smith, R., and Johansen, T.A. (1997), *Multiple Model Approaches to Modelling and Control*, London: Taylor and Francis.

Myers, R.H., and Montgomery, D.C. (1995), *Response Surface Methodology: Process and Product Design using Designed Experimetns*, New York: Wiley.

Neal, R.M. (1996), *Bayesian Learning for Neural Networks*, New York: Springer-Verlag.

Nishii, R. (1984), ''Asymptotic Properties of Criteria for selection of Variables in Mutiple Regression,'' *The Annals of Statistics*, 12(2), 758–765.

Nunnari, G., Dorling, S., Schlink, U., Cawled, G., Foxall, R., and Chatterton, T. (2004), ''Modelling SO2 Concentration at a Point with Statistical Approaches,'' *Environmental Modelling & Software*, 19(10), 887–905.

Orr, M.J.L. (1995), ''Regularisation in the Selection of Radial Basis Function Centers,'' *Neural Computation*, 7(3), 954–975.

Park, D.C., El-Sharkawi, M.A., and Marks, R.J. II (1991), ''An Adaptively Trained Neural Network,'' *IEEE Transactions on Neural Networks*, 2(3), 334–345.

Parlitz, U., Hornstein, A., Engster, D., Al-Bender, F., Lampaert, V., Tjahjowidodo, T., Fassois, S.D., Rizos, D., Wong, C.X., Worden, K., and Manso, G. (2004), ''Identification of Pre-sliding Friction Dynamics,'' *Chaos*, 14(2), 420–430.

Pelckmans, K., Suykens, J.A.K., and De Moor, B. (2005), *Neurocomputing*, 64, 137–159.

Peng, H., Ozaki, T., Toyodac, Y., Shioyad, H., Nakanoe, K., Haggan-Ozakif, V., and Mori, M. (2004), ''Modelling and Prediction of NOx Emission in a Coal-fired Power Generation Plant,'' *Control Engineering Practice*, 12, 191–203.

Peng, J., Li, K., and Huang, D.S. (2006), ''A Hybrid forward Algorithm for RBF Neural Network Construction,'' *IEEE Transactions on Neural Networks*, 17(6), 1439–1451.

Platt, J. (1991), ''A Resource Allocating Network for Function interpolation,'' *Neural Computation*, 3(2), 213–225.

Platt, J. (1999), ''Fast Training of Support Vector Machines using Sequential Minimal Optimization,'' in *Advances in Kernel Methods – Support Vector Learning*, eds. C.J.C. Burges, B. Schölkopf, and A.J. Smola, Cambridge, MA: MIT Press, pp. 185–208.

Plutowski, M., and White, H. (1993), ''Selecting Concise Training Sets from Clean Data,'' *IEEE Transactions on Neural Netwoks*, 4, 305–318.

Powell, M.J.D. (1985), "Radial Basis Functions for Multivariable interpolation: A Review," in *Algorithms for Approximation*, eds. J.C. Mansonn, and M.G. Cox, Oxford: Oxford University Press, pp. 143–167.

Priestley, M.B. (1981), *Spectral Analysis and Time Series*, London, New York: Academic Press.

Raghavan, H., Gopaluni, R.B., Shah, S., Pakpahan, J., Patwardhan, R., and Robson, C. (2005), "Gray-box Identification of Dynamic Models for the Bleaching Operation in a Pulp Mill," *Journal of Process Control*, 15(4), 451–468.

Rao, G.P. (2006), "Identification of Continuous-time Systems," *IEE Proceedings – Control Theory Applications*, 153(2), 185–220.

Ribeiro, B. (2005), "Support Vector Machines for Quality Monitoring in a Plastic Injection Molding Process," *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 35(3), 401–410.

Rojas, C.R., Welsh, J.S., Goodwin, G.C., and Feuer, A. (2007), "Robust Optimal Experiment Design for System Identification," *Automatica*, 43, 993–1008.

Ruano, A.E. (2005), *Intelligent Control Systems using Computational Intelligence Techniques*, London: IEE Publishing.

Sanner, R., and Slotine, J. (1992), "Gaussian Networks for Direct Adaptive Control," *IEEE Transactions on Neural Networks*, 3(6), 837–863.

Sargantanis, I.G., and Karim, M.N. (2004), "Variable Structure NARX Models: Application to Dissolved-Oxygen Bioprocess," *AIChE Journal*, 45(9), 2034–2045.

Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., and Murthy, K.R.K. (2000), "Improvements to the SMO Algortihms for SVM Regression," *IEEE Transactions on Neural Networks*, 11(5), 1188–1193.

Sindelář, R., and Babuška, R. (2004), "Input Selection for Nonlinear Regression Models," *IEEE Transactions on Fuzzy Systems*, 12(5), 688–696.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P., Hjalmarsson, H., and Juditsky, A. (1995), "Nonlinear Black-box Modelling in System Identification: A Unified Overview," *Automatica*, 31(12), 1691–1724.

Smola, A.J. (1998), *Learning with Kernels*. Ph.D. thesis, Germany: Informatik der Technischen Universitate Berlin.

Smola, A.J., and Schölkopf, B. (1998), "A Tutorial on Support Vector Regression," *NeuroCOLT Technical Report NC-TR-98-030*, London: Royal Holloway College, University of. http://www.kernel-machines.org/tutorial.html

Söderström, T. (2006), "Errors-in-variables Methods in System Identification," *Proc. 14th IFAC Symposium on System Identification*, Australia: Newcastle.

Söderström, T., and Stoica, P. (1989), *System Identification*, New York: Prentice Hall.

Söderström, T., and Stoica, P. (1990), "On Covariance Function Tests Used in System Identification," *Automatica*, 26, 125–133.

Söderström, T., Van den Hof, P., Wahlberg, B., and Weiland, S. (2005), "Special Issue on Data-based Modelling and System Identification," *Automatica*, 41(3), 357–562.

Soumelidis, M.I., and Stobart, R.K. (2006), "Dynamic Modelling of Three-way Catalysts using Non-linear Identification Techniques," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 220(7), 595–605.

Stewart, P., Fleming, P.J., and MacKenzie, S.A. (2003), "Real-time Simulation and Control Design by the Response Surface Methodology and Design of Experiments," *International Journal of Systems Science*, 34(14–15), 837–850.

Stone, M. (1974), "Cross Validatory Choie and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Series B*, 36, 117–147.

Sugiyama, M., and Ogawa, H. (2001), "Incremental Projection Learning for Optimal Generalization," *Neural Networks*, 14(1), 67–78.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J. (2002), *Least Squares Support Vector Machines*, Singapore: World Scientific.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statistical Society. Series B*, 58(1), 267–288.

Tothill, R.W., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., van Laar, R.K., Waring, P.M., Zalcberg, J., Ward, R., Biankin, A.V., Sutherland, R.L., Henshall, S.M., Fong, K., Pollack, J.R., Bowtell, D.D.L., and Holloway, A.J. (2005), "An Expression-based Site of Origin Diagnostic Method Designed for Clinical Application to Cancer of Unknown Origin," *Cancer Research*, 65(10), 4031–4040.

Townsend, S., and Irwin, G.W. (2001), "Nonlinear Model Based Predictive Control using Multiple Local Models," in *Nonlinear Predictive Control: Theory and Practice*, eds. B. Kouvaritakis, and M. Cannon, London: IEE, Control Engineering. Book Series 61.

Tsang, K.M., and Chan, W.L. (2005), "Adaptive Control of Power Factor Correction Converter using Nonliear System Identification," *IEE Proceedings Electric Power Applications*, 152(3), 627–633.

Vanderbei, R.J. (1999), "LOQO: An Interior Point Code for Quadratic Programming," *Optimization Methods and Software*, 11(1–4), 451–484.

Vapnik, V. (1995), *The Nature of Statictical Learning Theory*, New York: Springer-Verlag.

Vapnik, V. (1998), *Statistical Learning Theory*, New York: J. Wiley.

Vemuri, A., Polycarpou, M., and Diakourtis, S. (1998), "Neural Network Based Fault Detection and Accommodation in Robotic Manipulators," *IEEE Transactions on Robotics and Automation*, 14(2), 342–348.

Wahba, G. (1990), "Spline Models for Observational Data," *Society for Industrial and Applied Mathematics*.

Wang, L., and Mendel, J.M. (1992), "Fuzzy Basis Functions, Universal Approximation, and Orthogonal Least-squares Learning," *IEEE Transactions on Neural Networks*, 5, 807–814.

Wilson, D.J.H., Irwin, G.W., and Lightbody, G. (1999), "RbF Principal Surfaces for Process Monitoring," *IEEE Trans on Neural Networks*, 10(6), 1424–1434.

Xia, X., and Li, K. (2007), ''A New Score Correlation analysis Multi-class Support Vector Machine for Microarray,'' in *2007 International Joint Conference on Neural Networks*, Florida, Orlendo.

Xie, N., and Leung, H. (2005), ''Blind Equalization using a Predictive Radial Basis Function Neural Network,'' *IEEE Transactions on Neural Networks*, 16(3), 709–720.

Yingwei, L., Sundararajan, N., and Saratchandran, P. (1998), ''Performance Evaluation of a Sequential Minimal Radial Basisfunction (RBF) Neural Network Learning Algorithm,'' *IEEE Transactions on Neural Networks*, 9(2), 308–318.

Young, P.C. (1984), *Recursive Estimation and Time Series Analysis*, Berlin: Springer-Verlag.

Zhang, Q. (1993), ''Using Wavelets Network in Nonparametric Estimation,'' *IEEE Transactions on Neural Networks*, 8(2), 1997.

Zhang, L.F., Zhu, Q.M., and Longden, A. (2007), ''A Set of Novel Correlation Tests for Nonlinear System Variables,'' *International Journal of Systems Science*, 38(1), 47–60.

Zheng, G.L., and Billings, S.A. (1996), ''Radial Basis Function Network Configuration using Mutual Information and the Orthogonal Least Squares Algorithm,'' *Neural Networks*, 9, 1619–1637.

Zheng, G.L., and Billings, S.A. (1999), ''Qualitative Validation and Generalization in Non-linear System Identification,'' *International Journal of Control*, 72(17), 1592–1608.

Zheng, G.L., and Billings, S.A. (2002), ''Effects of Overparameterisation in Nonlinear System Identification and Neural Networks,'' *International Journal of Systems Science*, 33(5), 331–349.

Zhong, Z.D., Zhu, X.J., and Cao, G.Y. (2006), ''Modeling a PEMFC by a Support Vector Machine,'' *Journal of Power Sources*, 160(1), 293–298.